Studies of the Human Transcriptome

Sami Kilpinen

Institute for Molecular Medicine Finland
Faculty of Medicine
University of Helsinki

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Medicine of the University of Helsinki, for public examination in Lecture Hall 3, Biomedicum Helsinki, on June 17th, 2011, at 12 noon.

Helsinki 2011

Supervised by
Director, Professor Olli Kallioniemi
Institute for Molecular Medicine Finland (FIMM)
University of Helsinki, Finland

Reviewed by
Mauno Vihinen
Professor of Bioinformatics
Institute of Biomedical Technology
University of Tampere, Finland

And

Päivi Onkamo
Adjunct Professor in Genetic Bioinformatics
Department of Biosciences
University of Helsinki, Finland

Official Opponent
Inge Jonassen
Professor of Bioinformatics
Department of Informatics and
Computational Biology Unit
University of Bergen, Norway

ISBN 978-952-92-9105-2 (paperpack)
ISBN 978-952-10-7011-2 (pdf)
http://ethesis.helsinki.fi
Helsinki University Print 2011

“We are drowning in information and starving for knowledge.”
Rutherford D. Roger

1. List of Original Publications

This thesis is based on the following publications (referred by their Roman numerals I-IV):

I. **Kilpinen S***, Autio R*, Ojala K, Iljin K, Bucher E, Sara H, Pisto T, Saarela M, Skotheim RI, Björkman M, Mpindi J-P, Haapa-Paananen S, Vainio P, Edgren H, Wolf M, Astola J, Nees M, Hautaniemi S, Kallioniemi O. (2008) Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. Genome Biol 9: R139.

II. Autio R, **Kilpinen S**, Saarela M, Kallioniemi O, Hautaniemi S, Astola J. (2009) Comparison of Affymetrix data normalization methods using 6,926 experiments across five array generations. BMC Bioinformatics 10 Suppl 1: S24.

III. **Kilpinen S**, Ojala K, and Kallioniemi O. (2010) "Analysis of kinase gene expression patterns across 5681 human tissue samples reveals functional genomic taxonomy of the kinome." PLoS One **5**(12): e15068.

IV. **Kilpinen S**, Ojala K, Kallioniemi O. (2011) Alignment of gene expression profiles from test samples against a reference database: New method for context-specific interpretation of microarray data. BioData Mining Mar 31;4(1):5.

* Equal contribution
Manuscript of the publication I has been part of another thesis (ISBN 978-952-15-2029-7).

2. Abbreviations

AGC	Array Gene Centering
AGEP	Alignment of gene expression profiles
AML	Acute myeloid leukemia
AvgDiff	Average difference
BLAST	Basic Local Alignment Search Tool
CDF-file	Chip description file, describing the physical layout as well as as the probeset grouping of Affymetrix arrays
cDNA	Complementary DNA, usually DNA copy of mRNA.
CNS	Central nervous system
EQ	Equalization transformation
EST	Expressed sequence tags
GEO	Gene Expression Omnibus
GIST	Gastrointestinal stromal tumor
GO-BP	Gene ontology – biological processes
HK	Housekeeping gene based normalization
ICD10	International classification of diseases (version 10)
ICDO	International classification of diseases for oncology
IHC	Immunohistochemistry
IQR	Interquartile range
KL	Kullback-Leibler (distance)
LOOCV	Leave-one-out cross-validation
M-stage	Stage of tumor in terms of distant metastasis
MAQC	Microarray Quality Consortium
MAS5	Microarray Suite 5.0
MDS	Multidimensional scaling
MIAME	Minimum information about a microarray experiment
miRNA	Short ribonucleic acid (RNA) molecule
MM	Mismatch (probe)
mRNA	Messenger ribonucleic acid (RNA)
N-stage	Stage of tumor in terms of invasion to lymph nodes
NN	Nearest neighbour
PM	Perfect match (probe)
RMA	Robust multichip average
RNA-Seq	Ribonucleic acid sequencing
RT-PCR	Reverse transcription-polymerase chain reaction
SAGE	Serial analysis of gene expression
SMD	Stanford Microarray Database
SVM	Support vector machine
T-stage	Stage of tumor in terms of size and invasion status to nearby tissues
TM-score	Tissue match score
TS-score	Tissue specificity score
TTD	Therapeutic Target Database
WBL	Weibull distribution based normalization
Z	Standardized value

3. Abstract

Gene expression is one of the most critical factors influencing the phenotype of a cell. As a result of several technological advances, measuring gene expression levels has become one of the most common molecular biological measurements to study the behaviour of cells. The scientific community has produced enormous and constantly increasing collection of gene expression data from various human cells both from healthy and pathological conditions. However, while each of these studies is informative and enlighting in its own context and research setup, diverging methods and terminologies make it very challenging to integrate existing gene expression data to a more comprehensive view of human transcriptome function. On the other hand, bioinformatic science advances only through data integration and synthesis. The aim of this study was to develop biological and mathematical methods to overcome these challenges and to construct an integrated database of human transcriptome as well as to demonstrate its usage.

Methods developed in this study can be divided in two distinct parts. First, the biological and medical annotation of the existing gene expression measurements needed to be encoded by systematic vocabularies. There was no single existing biomedical ontology or vocabulary suitable for this purpose. Thus, new annotation terminology was developed as a part of this work. Second part was to develop mathematical methods correcting the noise and systematic differences/errors in the data caused by various array generations. Additionally, there was a need to develop suitable computational methods for sample collection and archiving, unique sample identification, database structures, data retrieval and visualization. Bioinformatic methods were developed to analyze gene expression levels and putative functional associations of human genes by using the integrated gene expression data. Also a method to interpret individual gene expression profiles across all the healthy and pathological tissues of the reference database was developed.

As a result of this work 9783 human gene expression samples measured by Affymetrix microarrays were integrated to form a unique human transcriptome resource – GeneSapiens. This makes it possible to analyse expression levels of 17330 genes across 175 types of healthy and pathological human tissues. Application of this resource to interpret individual gene expression measurements allowed identification of tissue of origin with 92.0% accuracy among 44 healthy tissue types. Systematic analysis of transcriptional activity levels of 459 kinase genes was performed across 44 healthy and 55 pathological tissue types and a genome wide analysis of kinase gene co-expression networks was done. This analysis revealed biologically and medically interesting data on putative kinase gene functions in health and disease. Finally, we developed a method for alignment of gene expression profiles (AGEP) to perform analysis for individual patient samples to pinpoint gene- and pathway-specific changes in the test sample in relation to the reference transcriptome database. We also showed how large-scale gene expression data resources can be used to quantitatively characterize changes in the transcriptomic program of differentiating stem cells.

Taken together, these studies indicate the power of systematic bioinformatic analyses to infer biological and medical insights from existing published datasets as well as to facilitate the interpretation of new molecular profiling data from individual patients.

4. Introduction

Since the sequencing of the human genome by Istrael *et al.*, Lander *et al.* and Venter *et al.* [1-3] there has been a rapid development of technologies enabling genome wide gene expression measurements as described by Schultze *et al.* [4]. The qualitatively and quantitatively increasing capability to analyze gene expression has greatly contributed towards our understanding of the functions of genes in health and disease.

One of the most widely applied technologies to perform genome wide gene expression measurements is the Affymetrix GeneChip system, which is based on 25mer probes that are photolithographically synthesized on the surface of the chip. During the many years of manufacture and application of these GeneChips, they have been found to be robust and reliable as described by Dalma-Weiszhausz *et al.* and Shi *et al.* [5-7]. Furthermore, the scientific community has greatly contributed to further development of data handling, normalization and data analysis methods available for Affymetrix based data with the few most influential studies being done by Bolstadt *et al.*, Faller *et al.*, Irizarry *et al.*, Schadt *et al.* and Workman *et al.* [8-12].

Constantly increasing amounts of gene expression data in public repositories as described by Edgar *et al.*, Hubble *et al.* and Rocca-Serra *et al.* [13-15] would allow for a much more advanced analysis of the molecular profiles of cells. Efforts to integrate these data together have been hindered by various technical difficulties resulting from incompatible microarray technologies and methods related to them. However, the scientific community has performed various meta-analysis studies of microarray data, with most influential studies done by Day *et al.*, Lee *et al.* and Rhodes *et al.* [16-20], partly overcoming these challenges by e.g. integrating the results obtained from the separate analysis of each of the studies.

The two most fundamental challenges are caused by the mutual incompatibility of various microarray generations and the heterogeneous anatomical, medical and pathological nomenclature applied to the annotation of the biological samples. It seems that in any gene expression measurement technology relying on hybridization of complementary nucleotide sequences, considerable variability is caused by the effect of the specific nucleotide sequence on the hybridization characteristics. This is a major reason for incompatibility between microarray generations. Any study attempting to combine and integrate data from multiple experiments needs to take this issue into account. Annotation of the biological samples needs to be sufficiently similar so that biologically and medically equivalent samples can be identified and grouped together for the purpose of data analysis. The enormous complexity of biological organisms and the range of clinical conditions renders most computational annotation methods ineffective and furthermore requires multiple layers of manual annotation to provide biologically sensible representation of samples.

Even though the layers of regulation of gene and protein expression and signaling in any cell are complicated beyond imagination, the mRNA

expression of genes is a key controller of cells' higher-level behaviour. Expression of genes provides components for the entire regulation machinery and expression level changes are required for fundamental changes in a cell's life. There is an enormous amount of scientific knowledge linking expression levels of various genes to myriads of phenomena in both healthy and diseased cells and tissues. However, most of these data remain scattered and fragmented, the synthesis and ultimate "model" of the human transcriptome is lacking.

However, if these challenges can be solved, the integration and synthesis of gene expression data provides novel possibilities to understand biological systems. For example, in order to identify a potential biomarker, one needs to be able to study the expression levels of the gene across the entire spectrum of tissues and diseases. Similarly, the association of an expression change of a gene to a specific disease or pathological state requires compatible expression data from both healthy and diseased tissues. Which genes change their expression levels in all epithelial malignancies when compared to all healthy epithelial tissues, can only be answered with an integrated data resource. The list of possible questions that can be answered is only limited by the amount of data accurately integrated. Thus, integration and synthesis of transcriptomics data is highly important for both the scientific understanding of cellular level phenomena as well as to support novel biomedical therapeutics and applications to heal diseases and manipulate biological organisms.

With the four studies presented here, we demonstrate one of the largest efforts to build a transcriptomic reference database and show its utilization to both studies of genes in health and disease as well as to interpret new gene expression profiles in the context of the reference database. In the first two studies we explain in detail how the database was constructed and validated, with a major advance achieved in solving the incompatibility between various Affymetrix array generations. In the third study, we systematically defined the expression level map of human kinase genes across major portion of healthy and diseased human tissues. We were also able to characterize functional associations of the kinase genes through the analysis of their genome-wide co-expression networks. The essential observation from this study was that kinase genes are indeed under unique transcriptional regulation so that accurate groups of pathological tissues (e.g. adenocarcinomas versus squamous) can be determined solely based on binarized kinase gene transcriptional profiles where the genes were divided into transcriptionally active and deactive states.

In the fourth study, we showed how large integrated reference database could be used to interpret expression profiles from individual new samples. One of the key advances of the study was a method enabling one to quantify similarity of individual expression profile against a reference database at the level of individual genes. This method has interesting applications, like the ability to interpret and compare expression profiles of patients against "peers" or the ability to identify and quantify changes in transcriptomic programs of differentiating cells.

5. Review of the literature

5.1. Transcriptome

The central theorem in classical biology is “DNA makes RNA makes protein”. Thus a cell transcribes various RNA molecules by using DNA as a template to be used as templates in translation. At any given time point, the collection of RNA molecules of a cell constitutes a transcriptome of a cell.

Methods and experiments aimed at measuring the expression values of single genes in dedicated experimental setups trace almost back to the origin of molecular biology. The limitations of the available methodologies forced these experiments to be focused on specific questions and the results were usually interpreted only in the context of certain cell lines, tissues or diseases. However, as early as 1999 Velculescu *et al.* [21] reported a study of the human transcriptomes of 19 normal and diseased human tissues by using serial analysis of gene expression (SAGE). Then in 2000, Warrington *et al.* [22] studied 11 different adult and fetal human tissues with high-density microarrays of that time to find out genes involved in cellular maintenance. They identified 535 genes from the studied 7000 genes as having a stable expression since it turns on during the fetal development. Additionally, they established average expression levels for genes in normal individuals and identified tissue specific genes for the 11 tissues. Later on Hsiao *et al.* [23] found 451 maintenance genes from 7000 studied genes to have a relatively stable expression across 19 distinct tissues while Eisenberg *et al.* [24] identified 575 maintenance genes from 7500 studied genes across 47 tissues. In 2002 Su *et al.* [25] studied 25 human and 45 mouse tissues, with a later study by Su *et al.* [26] containing 79 human and 61 mouse tissues. Also in 2005, Shyamsundar *et al.* [27] conducted a similar kind of study of gene expression in healthy human tissues reporting similarity in gene expression patterns between anatomically or functionally related tissues. Even though many of the early studies searched for the elusive maintenance, also known as housekeeping genes, they still represent the first attempts to characterize and understand human transcriptomes in a genome wide scale. In other words they were constructing first references against which other transcriptomic phenomena could be interpreted.

It was for long assumed that most transcribed RNAs are protein-coding, most likely directing the early development of methodologies towards enabling high-content analysis of these RNAs. Already in 2002 Kapranov *et al.* [28] found out that an order of magnitude more genomic sequence was transcribed than was accounted for by known or predicted exons. Practically all widely used microarray technologies still focus on polyadenylated RNA, which comprises only about 2% of the transcribed RNA molecules as described by Frith *et al.* [29]. The study of non-coding RNAs have been found to be a fruitful avenue as several studies revealed that non-coding miRNAs, cloned from *Caenorhabditis elegans* by Lee *et al.* as early as 1993 [30], were evolutionary widespread [31, 32] and have since shown to possess a wide variety of important regulatory functions [33]. These non-coding and non-polyadenylated RNAs were found to be an important transcriptomic regulatory mechanism [34] and within few years there were increasing

collection of registries, databases and tools for research around non-coding RNA molecules as described by Ambros *et al.* and Griffiths-Jones *et al.* [35-38]. The most recent research has revealed that the protein-coding part of the transcriptome is only a small portion of an otherwise extremely complex collection of various non-coding transcripts as revealed in studies by Frith *et al.*, Gingeras *et al.* and Kapranov *et al.* [29, 39, 40]. Strikingly, the ratio between non-coding and coding RNA molecules in a human transcriptome is 27:1, when excluding repetitive portion of the genome. According to Frith *et al.* [29] the ratio seems to increase with increasing complexity of an organism (1.1:1 in nematode, 2.2:1 in fruit fly and 28:1 in mouse). Actually, the entire concept of a gene is somewhat obsolete as the transcribed sequences are intertwined, nested and spliced in a complicated manner. During the past ten years this hidden part of the transcriptome has been brought to light, but the functions of those myriad transcripts remain rather unclear. What is clear, however, is that the complete transcriptome should be understood in much more complex terms with a very large number of distinct species of RNA molecules interacting in a highly complex manner to regulate the transcription and translation of protein-coding RNA species. Next-generation sequencing technology is rapidly combining these various research avenues as it allows an even more comprehensive analysis of transcribed RNA species as described by a series of recent studies by Metzker *et al.*, Mortazavi *et al.*, Pan *et al.* and Sultan *et al.* [41-45].

5.2. Gene expression analysis methods

There are numerous methods to measure the expression levels of one or more genes. The most well known methods are *in situ* hybridization, Northern blot and reverse transcription-polymerase chain reaction (RT-PCR). These are found to be robust and reliable, but they are somewhat limited in the number of genes (or samples) they can effectively measure simultaneously. However, early on there was a recognized need to measure the entire transcriptome at once, thus multiple methods were developed for that purpose. Expressed sequence tags (EST) were perhaps one of the earliest methods allowing genome wide analysis of gene expression. Later developed differential display, serial analysis of gene expression (SAGE), dot plots and nylon filter arrays and microarrays allowed more comprehensive genome wide expression measurements. The more recently developed RNA-Seq [41-45] allows perhaps the first true genome wide analysis of transcribed RNA-molecules. RNA-Seq is based on hugely parallel sequencing capacity allowing direct sequencing of RNA molecules from the sample. From these sequence reads one can then computationally form an estimate of expression levels of transcripts, their sequence variation and larger genomic rearrangements like fusion genes. Additionally, as the technology is not based on *a prior* assumption of transcript sequences it can also identify novel transcripts.

Microarray technology is currently the most established of these genome wide methods and is used widely in various research setups. Microarrays can be constructed with several methods like spotting (printing) cDNA sequences to glass slides with a robotic arrayer [46], by inkjet printer technology enabling noncontact printing by using electrical pulse to expel liquid to the glass slide [47] or by *in situ* synthesis [48, 49]. Most successful array manufacturers, like Affymetrix [7, 50] and Agilent Technologies [47] use *in situ* synthesis even

though the latter has also relied heavily on inkjet technology printing nucleotide by nucleotide to the glass slides (*in situ* printing). Illumina [51] has a somewhat differing concept where instead of using fixed positions for spots having oligonucleotide probes with specific sequences Illumina BeadArray technology synthesizes oligonucleotide probes on 3 μm silica beads which then self assemble in microwells. Affymetrix is by far the oldest and largest of the microarray manufacturers. This is also reflected in the amount of submitted microarray data in public repositories. For example, as of Dec 19, 2010, Gene Expression Omnibus (GEO) contained 90 827 Affymetrix based gene expression samples in comparison of 14 674 samples measured with Illumina and 4930 samples measured with Agilent Technologies. Next generation sequencing is rapidly replacing microarrays in almost all applications. However, the current availability of next generation sequencing data is nowhere near the amount of microarray data generated by the scientific community over the years.

In situ synthesis used by Affymetrix is based on light-directed synthesis of oligonucleotide probes on a silica substrate [49, 52]. The probes are built nucleotide by nucleotide by applying light on selected probes while the synthesis chemistry takes place only in the presence of light. A special mask is used to provide the exact configuration of light for each cycle of synthesis. Cycles are repeated until the desired probes are constructed.

There are two distinct types of gene expression microarrays in terms of sample hybridization protocol: single-channel and dual-channel arrays. The fundamental difference is that in dual-channel arrays two samples are differently labeled and hybridized onto a single array. The results are interpreted in terms of ratio between the different labels and thus reveal relative expression levels between the samples. In single-channel arrays only one labeled sample is hybridized to each array and therefore the results are interpreted more in a manner of absolute expression values than relative expression values. However, the requirement for the normalization of expression values results in non-absolute values even in the single-channel arrays. With single-channel arrays comparisons between the samples are done computationally. Two-channel arrays are somewhat outdated and most studies nowadays are done with single-channel arrays.

Affymetrix arrays are single channel arrays, consisting of 25 nucleotide long perfect match probes (PM) and mismatch probes (MM), together forming probe pairs. The perfect match probe is complementary to a desired position of specific RNA sequence while the mismatch probe is otherwise the same except the middle nucleotide is changed to complementary one. Ten to twenty of these probe pairs form a probeset. There are 6076 - 38191 probesets, depending on the array generation. About 80% of the probesets detect the antisense strand (mRNA) of the desired gene (these are denominated by "_at" at the end of the probeset ID according to Affymetrix probeset naming convention), about 10% cross-hybridize to same gene family (denominated by "_a_at"), about 5% cross-hybridize to some other gene (denominated by "__s_at") and about 5% contain at least one probe that hybridizes with some other sequence (denominated by "_x_at"). This setup of probes, probe pairs

and probesets has been subject to comprehensive review and improvement by the bioinformatic community as reviewed later on.

As Affymetrix arrays are single-channel only, one biotin-labeled RNA sample is hybridized on the array. The array is then stained with phycoerythrin-conjugated streptavidin, and after washing it is scanned with a Gene Array Scanner (manufactured by Affymetrix). The scanner provides the intensity values of each probe pair to be further processed by various algorithms.

There has been a lot of discussion about the quantitative accuracy of microarrays but Canales *et al.* [53] have shown that there is a good correlation between quantitative methods like RT-PCR and microarrays. Recently, comparison between Affymetrix arrays and RNA-Seq has also shown rather high correlations. Marioni *et al.* [54] showed that information in single lane of Illumina sequencing appears to be almost equivalent with single Affymetrix array in detecting differentially expressed genes. However, sequencing technology allows more sensitive detection of low expressed transcripts, alternative splicing and also is able to identify novel transcripts.

It has been known already for some time that fundamental limitations of microarray sensitivity exist especially in detecting low expression levels [55]. Similarly, according to Canales *et al.* [53] largest differences between the quantitative methods and microarrays are due to the lower sensitivity of microarrays at the low expression levels and due to the differing probe sequences. Also Hwang *et al.* [56], Nimgaonkar *et al.* [57] and by Autio *et al.* [58] revealed that differing probe sequences are a major source of noise when comparing expression level measurements from different technologies.

Through several studies, like the influential Microarray Quality Consortium (MAQC) [5, 6, 59], microarrays have been established as a robust and reliable technology to measure genome wide expression profiles. Especially Affymetrix arrays were found to have high reproducibility between laboratories [5] as well as to be reproducible between replicates according to Nimgaonkar *et al.* [57]. Barnes *et al.* and Jarvinen *et al.* [60, 61] report that there is considerable overall concordance between different microarray platforms. Recently microarray data has also shown to be in theory reliable enough for clinical use by the extension of the MAGC study (MAGC-II) [59]. However, the experimental setup and data-analysis quality of many studies leaves room for improvement before microarray-based classifiers can be used routinely in clinical practise. Nevertheless, some microarray-based tests, like MammaPrint [62], are already in clinical use.

5.3. Data processing and normalization of Affymetrix microarray data

Data produced by microarray scanners is considered to be raw data, as it requires substantial preprocessing and normalization before actual biological data analysis can be performed. Fundamental steps of preprocessing and normalization should contain at least a way to link the intensity of each measured probe (or probepair in the case of Affymetrix) to a preferentially

distinct biological feature like gene, transcript or exon. There should also be a way to deal with absurd intensity values likely resulting from technical artifacts as well as to compensate for variance in overall hybridization efficiency. The Affymetrix Microarray Suite version 5 (MAS5) [63] provides a suite of algorithms to perform the necessary preprocessing and normalization for Affymetrix arrays.

The scientific community has developed multiple additional data processing and normalization methods for Affymetrix arrays. In the same year as Affymetrix published MAS5, Li and Wong published the model-based expression index, dChip, providing another way to combine probe level intensity values into the final expression value of a probeset [64]. This was followed in 2003 by the highly influential Robust Multichip Average (RMA) by Irizarry *et al.* [65]. Several other methods like ChipMan, gMOS, GCRMA, PLIER, RSVD, UMTrMn, VSN, ZAM and ZL are expertly reviewed by Irizarry *et al.* [66]. Affymetrix arrays contain probes in pairwise manner, for each Perfect match probe designed to measure the transcript of interest the array contains also Mismatch probe. This latter probe is otherwise equal to Perfect match probe except that the middle nucleotide of the 25-mer is changed. Methods developed by the scientific community generally vary in how Perfect match and Mismatch probes are handled in the calculation of summary expression value and in the type of background correction made. However, irrespective of comprehensive studies of various preprosessing methods, like performed by Irizarry et al [66], there is no definite optimal preprocessing method for all purposes.

Affymetrix provides CDF-files containing array layout information, namely description of physical locations of the oligonucleotide probes on the array as well as information to which probeset each individual probe belongs to. In addition, Affymetrix provides information on which gene each probeset measures. This probeset to gene linking information can also be obtained from major genome browsers like Ensembl [67], UCSC genome browser [68] and NCBI genome browser [69]. However, Dai *et al.* [70] have shown that remapping of the probe sequences to the newer genome builds can significantly improve the data quality. Instead of relying on old definitions of which probe belongs to which probeset, Dai *et al.* [70] map individual probes to genes thus completely skipping the probeset level.

The need for normalization arises largely from the need to analyze multiple arrays together. In general, when one compares two or more arrays together one sees considerable variation in signal values. This variation can be broadly divided into biological one and technical one. Biological variation arises from the varying expression levels between the samples and it is usually the information that the researcher is seeking for. On the other hand, technical variation is, for example, caused by differences in sample handling, sample preparation or in the production of arrays or in the settings of the scanner. By far the largest challenge of the entire microarray field is to separate these two variations from each other and reliably eliminate technical variation. Affymetrix recommends a normalization where the total intensity of all probesets is scaled to be the same user-defined value across multiple arrays being compared together, but this simple approach does not perform well if

there are non-linear relationships between the arrays and practically not at all if there is a need for gene-specific normalization. This limitation arises from the fact that the scaling factor simply applies equal correction to all values within the array, thus it is unable to account for a need of any gene or value specific correction. RMA performs the widely applied quantile normalization, which replaces the maximum value of each array with the mean of maximum values; second largest value is replaced by the mean of the second largest values etc. This will give each array same distribution of values and is generally thought to be a relatively robust and efficient normalization. However, the approach has some challenges if all arrays do not have same set of genes and in the case of very large datasets there might be some computational challenges.

Normalizations generally applied to Affymetrix arrays, like scaling or quantile normalization, are suitable for reducing technical variation between arrays of the same generation. However, as Hwang *et al.* [56], Elo *et al.* [71], Canales *et al.* [53], Nimgaonkar *et al.* [57], Mecham *et al.* [72], Autio *et al.* [58] and many others have shown, one of the largest sources of noise in expression measurements originates from using nucleotide probes with varying sequences. This severely prohibits comparing or integrating data from multiple array generations. Hwang *et al.* [56] and Elo *et al.* [71] described methods how to improve comparability of Affymetrix array generations by selecting only a subset of probes. While this leads to a significant improvement in comparability it also greatly reduces the amount of usable data, as there is only a limited amount of overlapping sequences between the probes of two array generations. This is due to the logic of designing *in situ* synthesized oligonucleotide arrays, which leads to a complete redesign of probe sequences with new array generations with improved gene content. Other known approaches to perform cross-platform comparison include co-inertia analysis by Culhane *et al.* [73].

5.4. Sources of publicly available gene expression data

As the application of microarrays became widespread among the scientific community, the need for systematical storage of microarray results associated with publications increased in importance. To address this need, large bioinformatic projects were launched which resulted in construction of public expression data warehouses like Gene Expression Omnibus (GEO) [14], ArrayExpress [74] and Stanford Microarray Database (SMD) [15]. The primary aim of these warehouses was to enable systematic and long-term storage of large expression datasets and to allow retrieval of these datasets by the wide scientific community. For the sake of scientific credibility of these increasingly larger and more complex study setups, there was a need to describe the experiments in great detail. Brazma *et al.* [75] responded to this need and in 2001 published a standard known as minimum information about a microarray experiment (MIAME). Later array warehouses have mostly implemented this guideline and to some extent the details of the experimental setups of array studies have started to be more systematically described. In addition to these public gene expression data warehouses there is a large amount of gene expression data available at the websites of institutes, laboratories and research groups.

However, irrespective of the more strict guidelines and standards for publications, the actual repeatability of microarray-based studies is very low. This was demonstrated in a striking study by Ioannidis *et al.* where they showed that the data analysis of 10 out of 18 microarray based studies could not be reproduced based on their original publications [76]. The raw data itself produced by the modern microarray platforms is generally repeatable and reliable, but most publications relying on microarray-based data do not give adequate description of the used data analysis methods and their parameters nor do they release all data. This severely hinders real use of the results beyond the single publication. In many cases the only way for the scientific community to take advantage of and compare the results to other studies is to start with raw data and do the entire analysis again. However, this approach, when applied to multiple datasets, leads easily to the need for more and more complex microarray data meta-analysis methods and resources.

5.5. Meta-analyses of gene expression data

While public expression data warehouses like GEO [14], ArrayExpress [74] and SMD [15] served the main purpose of storing published data in a systematical manner, they did not originally support any analysis of the stored data. However, already 2003 Huminiecki *et al.* [77] showed that knowledge mining from large public databases of gene expression information can provide novel insights. One of their main results was that expression profiles extracted from variety of different sources of expression data (like Gene Expression Atlas [25], SAGEmap [78] and TissueInfo [79]) have relatively good correlations.

Once the meta-analysis of multiple datasets was shown to be a fruitful research direction by multiple authors [17, 77, 80-84], several gene expression databases and resources started to appear, with Oncomine [19], CELSIUS [16], Genevestigator [85] and BioGPS [86] being the most notable. These approaches have proven to be enormously useful as everyday genomics research tools. However, the biological heterogeneity inherent in all samples from biological organisms sets high requirements for the annotation of data in these reference databases. At present, computational text mining is not accurate enough to be able to handle biological complexity of the annotation even with the microarray experiment standardization efforts like MIAME [75]. Likewise, the data-driven computational approaches adopted by CELSIUS [16], where some of the biological characteristics of new samples are derived from the clustering of the samples among existing samples, is not able to handle the full biological complexity of sample annotation.

In addition to the biological challenges of the annotation, the mathematical challenges of data comparability also affect how meta-analysis studies are done. As demonstrated by Hwang *et al.* [56], Elo *et al.* [71], Autio *et al.* [58] there is a lot of technical variation between even the array generations of a single manufacturer (like Affymetrix Inc.), due to the different probe sequences. Previous correction methodology suggested by the same authors leads to the exclusion of incompatible probes and while it greatly improves the data comparability it also greatly reduces the amount of data. One might assume it to be the main reason why none of the large array meta-analysis studies have adopted those correction methods.

One of the largest and most influential meta-analysis projects done by the group of Arul Chinnayan, Oncomine [19], chose to represent its data study by study, thereby circumventing the comparability issue at the expense of data integration. In their original publication [19], they showed one of the first gene centric analyses by visualizing receptor tyrosine-protein kinase erbB-2 *(ERBB2)* gene expression levels across multiple tissue and then across multiple samples of healthy and ductal carcinoma of breast. This combined data from multiple datasets and allowed one to draw conclusions about the expression level activity of the *ERBB2* across various tissues. Also based on the hypothesis that therapeutic agents are most effective in cancers in which their targets are highly expressed they conducted a test of drug repositioning. By using Therapeutic Target Database (TTD) [87] and PubMed they identified 148 drugs and their targets. Then they proceeded to test in which cancers the target of the drug is statistically significantly overexpressed when compared to corresponding healthy tissue and discovered numerous interesting observations. They also published more advanced studies where they were able to show how genes with binding sites for typical cancer associated transcription factor like E2F were generally overexpressed in a variety of cancers whereas genes with a binding site for some other transcription factors like Myc-Max and C-Rel were overexpressed in specific types of cancers as described by Rhodes *et al.* [18]. This kind of analysis is an illustrative example of how gene expression meta-analysis can be used to uncover pathways related to the progression of cancer, a large mass of data leads to more reliable data analysis and allows more widely applicable conclusions to be drawn. The data integration approach chosen by Oncomine, while relatively simple approach to implement, leads to further challenges in development of data mining methods able to deal with fragmented datasets. Also visualizing various transcriptomic phenomena is challenging with fragmented datasets and therefore for a single question there might be multiple answers.

CELSIUS [16], Genevestigator [85] and many other projects have adopted practically the same approach. Higher numerical comparability has only been achieved in meta-analysis studies, like Greco *et al.*, Lee *et al.*, Segal *et al.* and Xu *et al.* [17, 80-83, 88], focusing on particular biological questions but not aiming to build integrated multiuse resource of transcriptomic data. On the commercial side GeneLogic Inc. aimed at building a comprehensive reference database and resolved the comparability issue by analyzing all relevant samples with a single array Affymetrix generation [89]. While this is undoubtedly the best approach, it is an economically completely unfeasible option in academic setting due to the constantly changing microarray platforms.

5.6. Gene expression – step between sequence and function

A gene's expression level provides intriguing information. Sequence level variation, or any causative link to the function of the protein encoded by the gene are hard or near impossible to derive from the gene expression data. The relation between gene expression and actual level of active protein is also hard or impossible to accurately derive from gene expression data. Nevertheless, it is one of the most important pieces of information, as the transcription of a

DNA sequence to mRNA is needed for translation and ultimately for the production of proteins, the functional components of cells. Therefore, since the sequencing of the human genome revealed a systematic catalogue of human genes, understanding the expression levels and ultimately functions of genes has become an ever more challenging and important task. The first step in understanding how the sequence transforms into function is to identify in which tissues genes are expressed. As previously described, there have been numerous studies establishing expression level information for an increasing gene and tissue content. As the expression levels of genes do not correlate in straightforward manner with the levels of proteins there should be separate analyses establishing both protein level and activity information. There are also highly succesfull studies establishing protein level information for majority of genes in considerably wide collection of tissues, such as the human protein atlas described by Uhlen *et al.* [90]. Newer technological advances allow higher content proteomics assays, such as lysate arrays [91], revealing protein levels across various healthy and diseased tissues in a single assay.

Cancer, a malignant neoplastic growth of a tissue, is a disease driven by various genetic changes and defects. Therefore the study of cancer-associated alterations, either at the level of DNA sequence changes, or at the level of gene expression is an important part of cancer research. Even though various sequence level changes of the non-transcribed parts of the genome might be indicative of or even causative for the disease, a key step in the understanding of the development of cancer is the analysis of the amounts of transcribed sequences and their exact sequence composition.

In the progression of cancer, one of the most studied families of genes is kinases. By phosphorylating various substrates kinases conduct and/or amplify signal transduction throughout the cell and therefore play an essential role in the signalling circuits of cells. Thus, for cancer cells, which reprogram various signalling circuits to enable their uncontrolled cell division and growth, kinase genes are especially critical. Indeed, among the known human cancer genes the most commonly represented protein domain is the protein kinase domain [92], indicating the essential role of kinases in the malignant progression. The most common cancer related genetic change targeting a kinase gene is an activating somatic mutation [92], but germ line mutations, recessive mutations, inactivating mutations, gene fusions, amplifications and deletions are also known. Some of these have an effect on the expression level of the kinase gene, like the amplification of *ERBB2* gene in ductal breast cancer leading to an overexpression of the transcript and subsequently to a larger amount of the corresponding tyrosine-kinase receptor at the surface of the cell [93, 94]. While kinase genes are among the most important genes to understand in the development of cancer, they are also very challenging to study for the reasons explained below.

For each kinase, it would be important to know i) what specific kinds of kinase gene sequences are transcribed ii) at what level they are transcribed iii) at what level the kinase proteins are present and iv) whether the kinase proteins are enzymatically active, and v) what is the actual biological function of the kinases. Even though there are some successful studies finding expression level signature indicative of the specific mutations [95, 96], direct sequence

analyses are needed to truly understand what is being transcribed. Next-generation RNA sequencing technology [41-45], allowing both efficient sequence analysis and expression level analysis from the same sample, is currently a promising way for the functional genetics advances.

Kinases have been the subject of various studies reporting sequence level changes in malignancies. Overall protein sequence similarity has been used to construct a sequence based classification of kinases [97, 98], comprehensive resources for studying kinase activity in various signalling pathways have been constructed [99] and various methods for defining the phosphorylation status of kinase substrates have been developed [100]. Nevertheless, due to the technical limitations kinase protein levels and enzymatic activity are practically impossible to measure across the entire kinome in all relevant tissues. Systematic expression level analysis focusing on kinase genes has been largely lacking, partly perhaps because of the difficulty of obtaining data for it and partly because kinases are mainly thought to function at the protein level without significant regulation at the transcriptome level.

Gene and ultimately protein sequence can be used to make inferences as to the function of the protein, like in the case of kinase genes the domain responsible for the kinase activity can be recognised from the sequence. Mutations in a specific nucleotide of the gene can be predicted to have effect on the function of the protein. However, this kind of analysis cannot always reveal higher-level biological processes in which the protein participates nor are the protein domains always conserved enough to be recognised. Genome-wide gene expression measurements not only provide the possibility to characterise which genes are expressed in which tissues, but also provide tentative information as to which biological functions the gene products might participate in. Merely finding a group of genes differentially expressed in a group of tissues or cells subject to a specific perturbation might indicate that corresponding genes are participating in certain function responding to the perturbation. Taking this simple assumption somewhat further leads one to the co-expression analysis where correlating expression levels between genes provide an indication of similar functions as previously studied by Lee *et al.*, Prifti *et al.* and Zhang *et al.* [17, 101, 102]. This is especially true in the case of protein complexes as the complex is rarely functional if all of its components are not present. It has been shown that if some of the co-expressing genes have known functions then under certain assumptions these functions can be assumed for unknown genes. These methods have been expertly reviewed by Hu *et al.* [103]. Segal *et al.* and Xu *et al.* [80-83] took the study of gene function even further by demonstrating how one can identify networks of interacting modules of coregulated genes. Methods used in these studies vary somewhat, but the core idea is the same. Having a large collection of integrated expression data makes it possible to uncover functional associations of genes through careful analysis of their co-expression environment.

5.7. Interpreting microarray data in the context of existing data

On the field of nucleotide and amino acid sequence analysis tools, like BLAST and BLAT [104, 105], enabling comparison of an unknown sequence to a reference database of known sequences has proven to be essential. Similarly on the side of gene expression data analysis interpreting new data in the context of existing data has been found to be a useful approach. One of the earliest to demonstrate this was Parmigiani *et al.* [106] in 2002 who constructed a statistical framework allowing probabilistic assignment of tumors to molecular profiles. This has been followed by many others like Zilliox *et al.* [107] with their gene expression barcode methodology predicting tissue type of individual sample with the help of gene expression barcodes constructed from a reference data. Lamb *et al.* [108] constructed Connectivity Map, showing how different drugs change the expression profile of various cell lines and enabled comparison of gene expression changes observed in one's own studies to the established expression changes caused by drugs. Caldas *et al.* [109] demonstrated a methodology to retrieve experiments resembling the one's own experiment based on the measured expression values. More recently Lopez *et al.* published [110] TranscriptomeBrowser, a resource allowing search of transcriptomic signatures from a large collection of microarray experiments.

The defining aspect of all of these is that similar experiments are not identified based on the similar annotation but based on similar data values. Therefore these can be seen as analogs of BLAST [105] and BLAT [104] types of sequence analysis tools where the sequence of an unknown sample is being compared with those of all other sequences available in the reference sequence databank (like GenBank). However, using a gene expression database as a reference to interpret new samples is somewhat more complicated than comparing a nucleotide or amino acid sequence to a database of sequences. Most importantly, there is no simple definition of similarity between expression level of a gene in the query sample and its expression in the reference database.

Nevertheless, the ability to compare an individual expression profile against reference data could provide new tools for personalized medicine. The scientific literature describing cancer related gene expression changes and signatures is rapidly increasing, but very few of those findings have been transferred into clinical practise. Reasons are numerous, but one specific challenge is that gene expression measurements from the patient's tumor itself are rather difficult to interpret without a proper reference. At 2006 Gruvberger-Saal *et al.* [111] expertly reviewed many of the challenges of using microarrays in clinical settings. They especially pointed out a need for standardization of the methods, arrays and easier comparability between the studies. It still remains to be seen how established microarray based diagnostics tests, like MammaPrint or TargetPrint [112], perform outside the patient population used to develop those tests. Quite often microarray based classifiers are validated with a limited population of samples, perhaps some specific ethnic group or disease subtype. Therefore it is obvious that more standardized reference data is needed in large quantities as well as methods to robustly compare patients to it.

6. Aims of the study

The aims of the study were to

- Collect a significant amount of published human gene expression data into a unified database and apply a systematic annotation to the samples.
- Develop methods to overcome mathematical and biological challenges in data integration across different microarray platforms.
- Develop methods to mine the integrated data both in a gene wise and sample wise manner in order to acquire new biological and biomedical knowledge.
- Develop methods and statistical tools to compare molecular profiling data from one sample against a comprehensive collection of annotated reference data.

7. Materials & Methods

Each publication (I-IV) describes in detail all the materials and methods used in it. However, the main methods are briefly previewed here for completeness and convenience for the reader.

7.1. Data acquisition and archiving

Data used in publication I was collected in the form of Affymetrix CEL files (containing intensity data as measured by the microarray scanner) mainly from public sources like GEO and ArrayExpress. Some additional studies were obtained directly from authors of substantial gene expression experiments. As any collection of CEL files might theoretically contain duplicate files the uniqueness of each CEL file was tested by using the cyclic redundancy check algorithm (cksum) [113]. Cksum provides "fingerprint" of the content of the file, usually used to check integrity of files, but it can be adapted for this purpose as well. This step significantly reduced the risk of including the same sample twice in the data collection. Data was archived in a Linux-system with additional Perl scripts to maintain the archive integrity and calculate the cksums.

7.2. Data integration

7.2.1. Data preprocessing

All CEL-files were preprocessed with the Microarray Suite 5.0 (MAS5) algorithm, implemented with C++ by using libraries provided by Affymetrix Inc. MAS5 produces both quantitative expression values as well as qualitative values from the raw data file (CEL-file). MAS5 performs a background correction by calculating and subtracting the weighted sum of the background signal of the various zones of the array from the values of the individual spots. A detection call, a qualitative value, indicates whether the transcript is reliably detected (Present) or not detected (Absent) by the probes of the array. However, the normalization schema used in this study does not use detection calls.

The quantitative expression value is calculated by One-step Tukey's Biweight Estimate to represent the level of expression of the corresponding transcript. First the signal of each probe pair is estimated with the log of Perfect match probe intensities after a subtracting stray signal estimate. The stray signal estimate is formed according to three rules as described in the Affymetrix statistical reference guide [63]

i) if the Mismatch probe intensity value is less than the Perfect match probe value, the mismatch intensity is considered to be informative and a proper estimate of the stray signal,
ii) if the Mismatch probes are generally informative across the probeset except for a few probes, the stray signal is calculated as the bi-weight mean of the Perfect match and Mismatch ratio,
iii) if the Mismatch probes are generally uninformative, the stray signal is defined to be slightly less than Perfect match signal.

The closer the signal of the probe pair is to the median of all probe pairs of the corresponding probeset the stronger the weight that probe pair gets. These weights are then used in the weighted mean of all probe pair signal values to determine the final signal value.

To avoid the somewhat obsolete probeset to gene mapping provided by Affymetrix Inc. we used alternative CDF files mapping individual probes directly to the Ensembl gene IDs. This is described in more detail both in section 5.3 and in publication I.

7.2.2. Samplewise normalization

Equalization transformation was used to normalize each sample preprocessed with MAS5. Mean 8 and standard deviation 2 were selected as parameters for desired distribution based on the comparison to median (7.92) and standard deviation (2.3) of all 9783 samples. Equalization transformation is described in more detail in publications I and II as well as by Hautaniemi *et al.* [114].

7.2.3. Genewise normalization

Array-generation-based gene centering (AGC) was performed to alleviate noise from varying probe sequences between array generations. In AGC each gene is corrected for array-generation-based bias of measuring the expression. This is based on the assumption that having a large enough collection of samples analyzed the distribution of values of a gene contains all possible expression values across all tissues for each array generation. Thus the difference between the distributions of the gene between array generations is largely due to the technical variation caused by varying probe sequences. AGC is described in more detail in publications I and II.

7.3. Data annotation

7.3.1. Sample and gene annotation

Annotation provided by the original authors was retrieved with Perl scripts (for GEO and ArrayExpress data) or manually from publications and their supplementary tables. The annotation provided information about the biological nature of the sample. A team of biologists and medical doctors then manually curated annotation of each sample resulting in 17 fields of information relevant to samples biological characteristics. The information includes for example anatomical system from which the sample originates, pathological status, sex and age. These fields are listed in Table 1. This annotation was regarded as primary annotation.

A secondary annotation layer was then constructed by defining groups of samples having certain combinations of primary annotation values. This allowed easy implementation of different levels of annotation, like sample group “breast cancers” and separate sample groups of histological subtypes of breast cancer.

As the gene definition in GeneSapiens is based on Ensembl genes the annotation for each gene was fetched from Ensembl by using custom written Perl scripts.

7.4. Data validation

7.4.1. Multidimensional scaling

In publication I classical multidimensional scaling [115] was used to visualize distances between thousands of gene expression profiles to understand what kind of clusters they form. Manhattan distance, also known as taxicab distance, was used as distance metric (see Equation 1 for Manhattan distance between two points in two-dimensional space).

Equation 1

$$d((x_1,y_1),(x_2,y_2)) = |x_1 - x_2| + |y_1 - y_2|$$

Calculation of the Manhattan distance between all pairs of samples results in a symmetrical distance matrix of thousands of rows and columns. Multidimensional scaling (MDS) was used to reduce the number of dimensions of the matrix and to represent the distances between samples in a selected number of dimensions. In publication I, three-dimensional projection was used whereas in Figure 6 two-dimensional projection was used. Additionally, in Figure 6 Pearson correlation coefficient was used as the distance metric. In MDS figures the distance between dots, each representing an individual sample, approximate the true (Manhattan or Pearson correlation) distance between the expression profiles.

7.4.2. K-means clustering and rand index

In publication I k-means clustering was used to test the goodness of the normalization. K-means clustering can be used to partition gene expression samples into k clusters. K-means clustering assigns each sample to a cluster whose mean expression profile is most similar to the sample. Clustering was performed with default parameters in R and allowed to run a maximum of 100,000 iterations and the initial centres of the clusters were given as median profiles of either array generations or tissues. The aim was to test whether samples form clusters based on their array generation type or rather based on their tissue type. The results of k-means clustering was tested with corrected rand index [116] by using the flexible procedures for clustering (fpc) library in R. The corrected rand index can be used to quantify how randomly class labels are assigned into different clusters. The corrected rand index varies between 0 and 1. A 0 means that class labels are randomly segregated among the clusters and 1 means non-random segregation (like all occurrences of one class are in single cluster).

7.4.3. Kullback-Leibler divergence of housekeeping genes

Housekeeping gene expression stability was one measure of the goodness of normalization used in the publication II. It was assumed that expression values of each housekeeping gene are distributed similarly across all the array generations. To measure the differences in distributions we used the Kullback-Leibler divergence (KL-divergence) [117, 118]. This measure of divergence can be used to quantify how much two distributions differ. The range of expression values of each housekeeping gene was divided into 50 bins so that each bin contains 2% of the expression values. Then the distribution of expression values of housekeeping genes across all array generations was

compared with KL-divergence to distribution of expression values in each array generation. We used 126 housekeeping genes [24] in this analysis. The aim was to show that the KL-divergence between housekeeping gene expression value distributions in each array generation and the combined distribution across all array generations diminish after the normalization. This indicates that after the normalization housekeeping gene expression profiles are more alike irrespective of the array generation type used to measure them and therefore the data is biologically more sensible.

7.5. Data analysis methods

7.5.1. Definition of transcriptional activity

In publication III we demonstrated a method to define the state of transcriptional activity of genes. The method relied on defining background expression level by calculating the entropy of sample annotation class labels in a sliding window (with width of 5% of maximum expression value of the gene in question) over 1603 healthy tissue samples. Background expression level was defined to be smaller or equal than the midpoint of the sliding window with highest entropy value plus two times the standard deviation below the midpoint. This results in 95% coverage of the assumed normal distribution of the background expression level of the gene. If the median expression of gene in a tissue was above the background expression level it was defined as transcriptionally active and assumed to be under active and positive regulation of transcription in that tissue.

In the analysis of all human genes (Figure 8, Figure 9 and Figure 10) the method was further modified to more effectively handle genes actively transcribed in all tissues (i.e. housekeeping genes). The location of the sliding window with the highest entropy was further analyzed to identify housekeeping genes. If the midpoint of the sliding window was above the lower 35:th percentile of expression values and the sum of all entropy values at lower expression values than the location of the sliding window was less than 0.95 of the theoretical maximum entropy sum, the gene was classified as housekeeping gene. After this modification the method is able to classify genes being highly expressed in all samples as housekeeping genes.

7.5.2. Co-expression environment biological process enrichments

Correlation coefficients between all gene pairs (11 906 genes) in the genome were calculated over 5712 samples, and the genomic co-expression network was constructed by forming links between genes whose correlation coefficient was over the gene specific threshold. Threshold was defined to be the 99.9 percentile of all correlations for the gene in question. This gene specific threshold was found to be important factor as each gene seems to have a different range of correlation coefficients with other genes. Then the co-expression network around each of the kinase genes was explored by performing 500 random walks on the network, each 5 steps long and originating from the gene of interest, collecting all other genes (nodes of the network) encountered. Steps were not allowed to go directly backwards. As a result of this we acquired the frequency distribution of genes encountered in the near vicinity of the kinase gene. This distribution of genes was then

analysed in terms of the relative enrichments of biological processes as defined by Gene Ontology (GO-BP) by using the GOSim R library [119]. The method took into account both the topology of the genomic co-expression network as well as the dynamic range of correlation coefficients each gene has.

7.5.3. Calculation of gene expression density estimates

For the purpose of alignment of expression profiles to the reference data we calculated density estimates of the expression distribution of each gene in each reference tissue. Simplified schema of the density estimate formation is shown in the Figure 1. The estimates were calculated by using fast Fourier transformation based approximation of kernel density estimates with Gaussian window. Bandwidth selection was done as described by Scott *et al.* [120]. The density was estimated from 0 to maximum expression value in the entire dataset plus two times the highest bandwidth for that gene, with 512 equally spaced points.

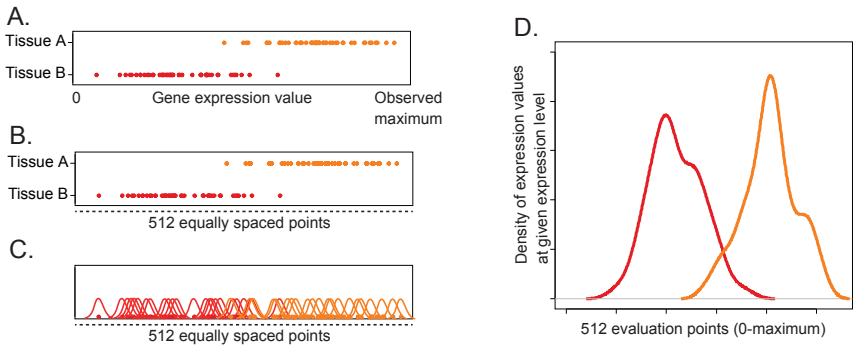

Figure 1 Simplified schema of forming expression density estimates of a gene in two tissues. A) Individual expression measurements of a gene in two tissues. B) The range between 0 and maximum expression value of the gene is divided into 512 equally spaced points. C) One way to visualize density estimate formation is to situate normal distribution around each datapoint with a mean value of the value of datapoint and standard deviation equal to the bandwidth. D) Final density estimate can then be understood as a sum of normal distribution values at each evaluation point. The resulting diagram has an area of one and the height of the diagram at each evaluation point indicates the density of observations at that point. In other words, density estimate describe what expression levels are typical for the gene (in the tissue in question).

7.5.4. Alignment of an external sample to gene expression density estimate

Constructing gene and tissue specific density estimates of a reference database allows one to compare a single external expression profile to the database. Expression values of each gene in the external query sample are compared to density estimates of the gene in each reference tissue. Separately for all tissues, the closest evaluation point is identified and corresponding density value is found. The “goodness of match” of the query sample to the reference tissue in terms of the particular gene is the proportion of evaluation points having a lower density value than the gene’s value in the query sample. The resulting value is called the tissue match score (tm-score) and it varies between 0 (not “matching”) and 1 (perfect “match”). For example, if an expression value of the query sample for a gene situates at the evaluation point having the highest density in the tissue then the gene gets tm-score of 1 against that tissue as it is the best “matching” expression value for the gene in

that tissue. This is repeated for expression values of all genes in the query sample, essentially leading to a matrix of tm-scores (tm-score for each gene in each reference tissue).

Even though the tm-score defines how well the expression value of a gene in the query sample matches the expected expression values of the reference tissues it does not tell whether the expression value is at a level unique for the reference tissue. To find how uniquely expression values of the query sample matched each tissue we transformed tm-scores into tissue-specificity scores (ts-scores). The ts-score for a gene for a tissue is the mean of the ratio-weighted differences of the tm-score for the tissue and tm-scores for all other tissues. The ratio weighting is done so that the larger the ratio between the tm-scores, the higher the resulting ts-score will be. Ts-score indicates how well the tm-score of a gene categorizes the query sample into a tissue. The ts-score varies between -1 and 1. A ts-score of 1 for a gene indicates that the query sample had an expression value specific for the tissue in question. A ts-score of -1 indicates that the gene has a specific expression value for the tissue in question but the query sample did not have that expression value. The mean of the ts-scores for a tissue is used as a score describing the similarity of the query sample to the tissue.

7.6. Visualization methods

7.6.1. Body-wide expression profiles of genes

To provide a comprehensive view of gene expression over healthy and pathological tissues specific visualization methods were developed. In these body-wide expression profiles the expression level of the gene is on the y-axis. On the x-axis are all samples of the database ordered in terms of the anatomical origin of the sample as well as in terms of the pathological status of the sample, thus each dot represents the expression level of a gene in one sample. The anatomical origins of each sample are marked with colored bars below the plot. In the first part of the plot are healthy tissue samples, in the second part malignant tissues and in third samples of other diseases. Tissues having an expression level at least one standard deviation higher than average expression of all tissues of the same type (healthy, cancer, or other disease) or ones whose 90th percentile of expression in a tissue is equal or higher than 2*interquartile range (IQR) + 75th percentile of the same type (healthy, cancer, or other disease) are additionally colored in the figure (legend at the top left corner of the image). These plots have been used in publications I, III and IV.

Another frequently used plot type is boxplot. On the left side of these plots are healthy tissues (green) and on the right are malignant tissues (red). The number of samples per tissue is shown in the parentheses. The line shows the median expression level. The box extends to the 25 and 75 percentiles with whiskers extending to the most extreme data point which is no more than 1.5*IQR from the median. Outliers beyond this are shown as individual dots. These plots have been used in publications I, III and IV.

7.6.2. Visualization of co-expression data

In publication III the co-expression enrichments of kinase genes were visualized in a heatmap format. On the x-axis were kinase genes and on the y-axis GO-BP classes. The x-axis was clustered by using binary distance and complete linkage whereas the y-axis was clustered by using Lin semantic similarity metric [121] and Ward linkage metric. Black color indicates that in the co-expression network of the kinase gene in question the GO-BP class in question is statistically significantly enriched. Individual GO-BP classes were grouped under a representative name for the sake of readability.

7.6.3. Body-wide gene expression heatmaps

Similarly to provide an overview of the expression of an entire gene family across healthy and pathological tissues specific style of heatmap visualization was developed. This bodymap contains the mean expression values of chosen genes (x-axis) in 111 human *in vivo* tissues (y-axis). The cells of the map have been colored so that red indicates high expression and blue low expression. The mean values of the genes were centered by subtracting mean for the gene and scaled by dividing with the root-mean-square. The number of samples available for each tissue is shown in the parentheses. Both axes of the map have been clustered (Euclidean distance with ward linkage) so that tissues resembling each other the most are closest to each other, and similarly, genes with most similar expression patterns are placed adjacent to each other. These plots have been used in publication I. Differently to that, in publication III we used a heatmap with binarized expression levels. There clustering was based on binary distance and complete linkage The complete linkage tree building algorithm used in publication III provides larger clusters and is more suitable for generating an overall view of the broad groups of the kinome transcriptome.

8. Results

8.1. Constructing GeneSapiens

We collected 9783 Affymetrix CEL files from public sources, which were then used to construct a synthesis of genomewide expression profiles across 175 different healthy and pathological tissues. As same CEL files might have been used in multiple studies their uniqueness was tested by using cyclic redundancy check algorithm. Further steps in the construction of the database can be divided into data integration and annotation.

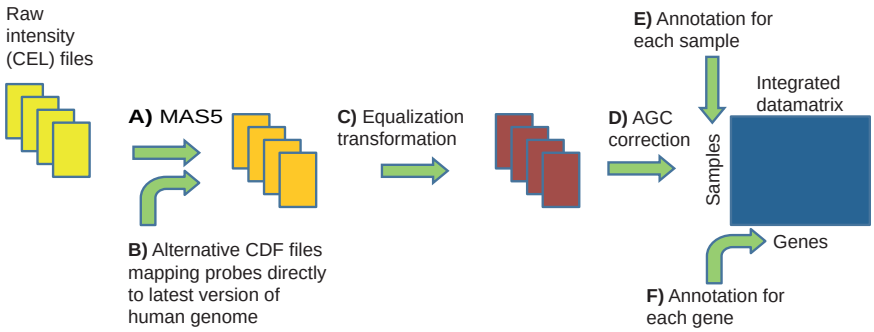

Figure 2 Schematics of constructing integrated Affymetrix gene expression dataset. A) MAS5 is used as a preprocessing method performing both background correction and combining probe level data into a gene level data (with alternative CDF-files), B) Alternative CDF-files are used to link probe level data directly into a gene level data, C) Equalization transformation is used to perform samplewise normalization, D) Array gene centering (AGC) is used to perform genewise normalization across array generations, E) Each sample is then linked with curated annotation of biological details of the hybridized sample, F) Each gene is further annotated from suitable resource (like Ensembl [122]).

8.1.1. Data integration

As a first step in data integration, raw data from CEL files was preprocessed using MAS5. Alternative CDF files were used to map probes directly to Ensembl gene ids (Figure 2A-B). When using alternative CDF-files all the probes identifying a gene with high enough specificity effectively form one large probeset [70]. The usage of these files provide significant benefits as a more up to date genome build is taken into account when defining which gene each probe measures, and ultimately, each gene can be assigned with only one expression value in each sample.

As shown by Canales *et al.* [53], Nimgaonkar *et al.* [57], Hwang *et al.* [56] and Autio *et al.* [58] varying probe sequences (Figure 3A) is one of the main sources of technical variation between array generations. It seems that a change of even a single nucleotide affects the hybridization efficiency of the probe enough to cause a noticeable change in the final expression value. This level of sensitivity is usefull to calculate stray signal estimate from Mismatch probe data. However, even completely complementary probes for a gene can be major source of noise if they measure different parts of the sequence of the gene. For example, when measuring a single mRNA sample with two different array generations (HG-U95 and HG-U133A) the correlation coefficient between expression values might be as low as 0.52. However, if one filters

probes based on the sequence overlap (between the probes of the array generations) the correlation coefficient systematically increases (Figure 3B). While this approach provides viable solution for incomparability of results between different array generations it remains impractical as the number of genes remaining in the analysis rapidly decreases (Figure 3B). The reasons for this probe sequence dependent noise are multiple. For certain, a part is caused by different hybridization kinetics of different nucleotide sequences, but the issue of alternative splicing is also involved. Various array generations might actually measure different splice variants of the genes and therefore produce different expression values. Irrespective of the reasons, the probe sequence dependent noise might have a wider impact on various molecular genetic measurement technologies than is generally known.

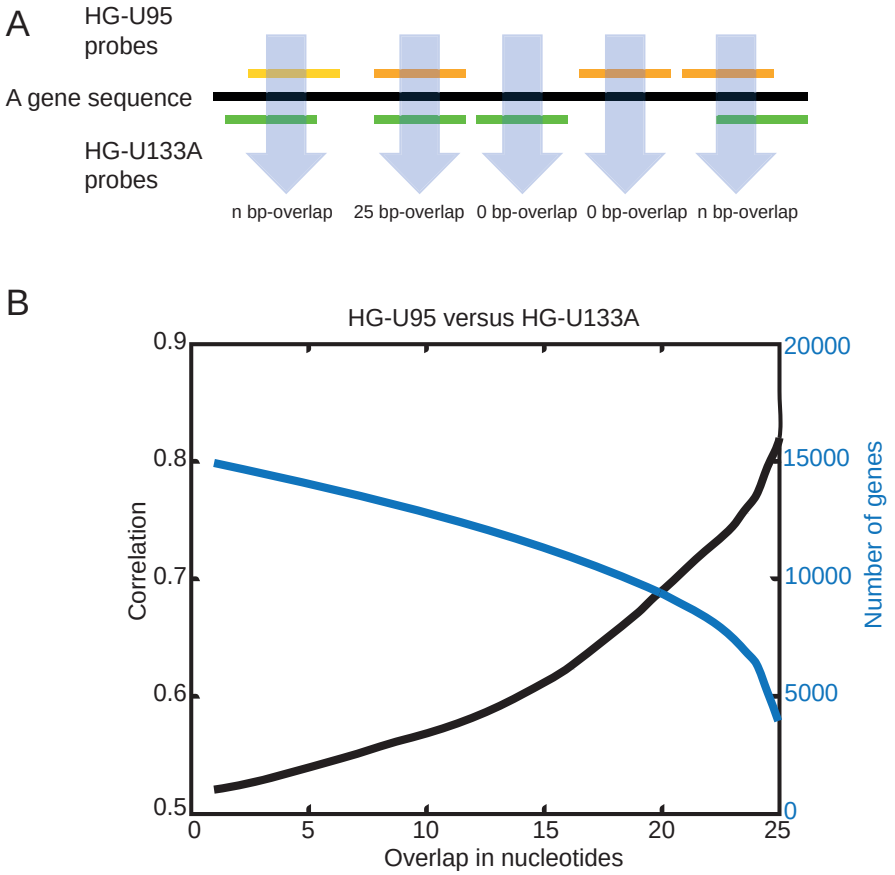

Figure 3 A) Conceptual drawing of four different oligonucleotide probes from two different Affymetrix array generations. Each probe hybridizes to a different part of the sequence of the gene and therefore has varying amounts (0-25 bp) of overlapping nucleotide sequences with the probes of another array generation. B) The black line depicts the effect of varying probe sequence overlap on correlation between replicate analysis of same RNA sample on two array generations, HG-U95 and HG-U133. The blue line depicts the number of genes remaining in the analysis after filtering the probes based on nucleotide sequence overlap between array generations. Figure produced from the data described in Autio *et al.* [58].

This approach of probe filtering based on sequence overlap has been utilized in one form or another in a few array meta-analysis studies [56, 71] but as our aim was to build a comprehensive transcriptomic resource by using data from multiple array generations the probe filtering approach was not viable option, mainly due to the impossibility of finding a reasonable set of overlapping probes between several array generations. For example, to achieve correlation coeffiencient of 0.8 between replicates one would need to use only probes with complete 25 bp overlap and that would drop the amount of genes to less than 5000. Additionally, the expression level of each gene would be derived from significantly smaller number of probes, thus the noise can be assumed to increase substantially.

Instead we studied another approach utilizing the large data collections to correct for probe level noise. In theory, one can assume each gene to have certain biologically meaningful range of expression values, a certain minimum background expression level and a certain maximum level where it is expressed as actively as possible. Then, if one has a large enough collection of genome wide expression data one can assume to have this biologically meaningful range of expression values covered. In other words, with a certain accuracy one might assume the distribution of expression values of each gene to cover all biologically sensible expression levels (states). Then, given this amount of data from each array generation one is able to identify a major portion of the noise caused by probe sequence effects by comparing the expression value distributions of genes between the array generations (Figure 4A-B). This method was named as Array Gene Centering (AGC) and we were able to show it to work well enough to perform genewise correction of noise caused by varying probe sequences. The normalization was considered to perform well enough if in normalized data biological variation clearly surpassed the technical variation. This would allow biological conclusions drawn from the data. Actually, it is reasonable to assume that AGC compensates for a variety of sources of noise when comparing different array generations, not only for probe sequence effects. As AGC is based on few simple assumptions about the distributions it should be applicable for normalizing other high-throughput data as well.

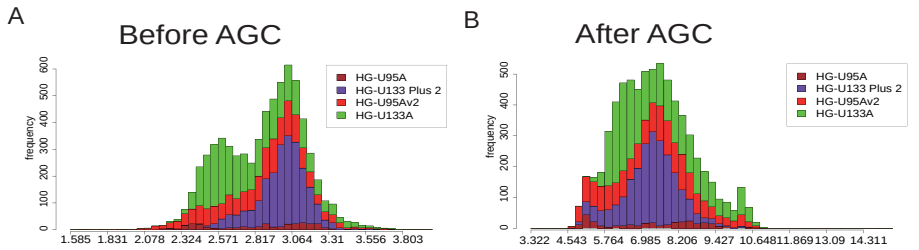

Figure 4 The effect of AGC correction to the data of a single gene. A) Histogram of expression value distributions of the gene across four array generations before AGC correction. B) Histogram of expression value distributions of the gene across four array generations after AGC correction.

The normalization of GeneSapiens data was two-dimensional (Figure 5A-B). Samplewise normalization with Equalization transformation (EQ) was done to correct for technical variation between samples (Figure 2C and Figure 5A). The parameters of the distribution to which data was normalized with EQ (mean=8, standard deviation=2) were selected based on the entire dataset of 9783 samples. Application of AGC as genewise correction for EQ normalized data (Figure 2D,Figure 5B) resulted in one final expression value per gene per sample, altogether 113 786 107 gene expression values. See materials & methods as well as publications I and II for more details on data integration.

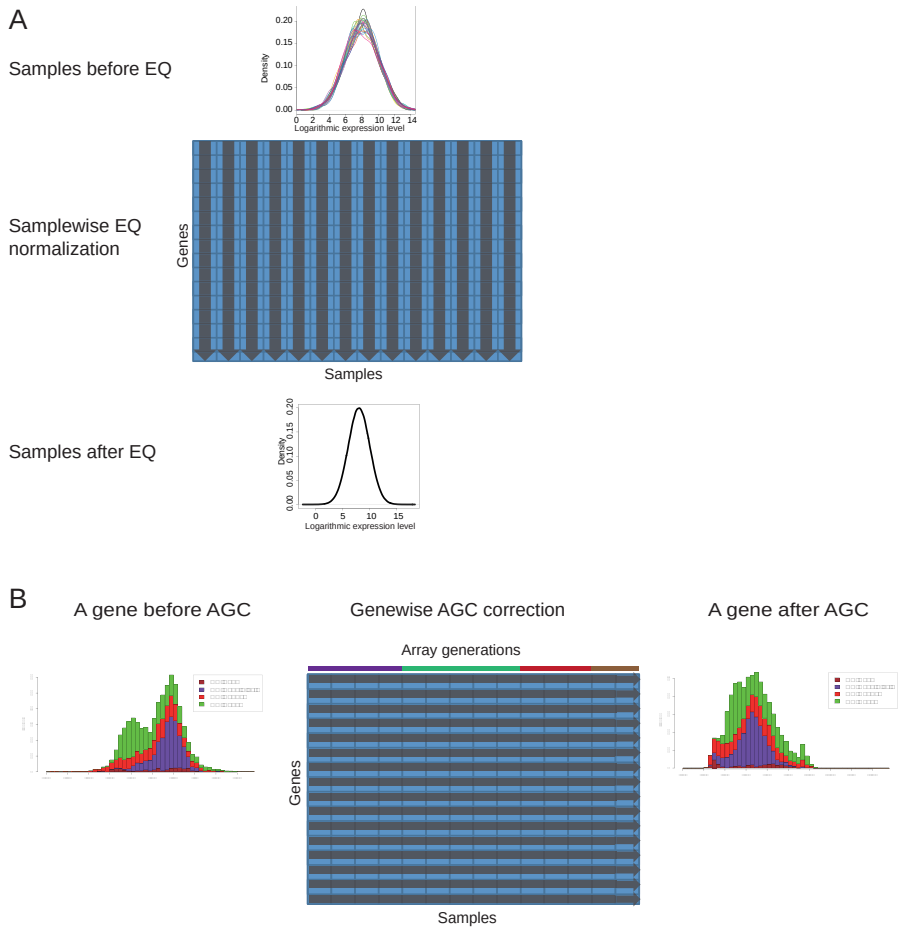

Figure 5 Schematics of two-dimensional normalization scheme used in the GeneSapiens. A) Each sample is normalized with equalization transformation (EQ) B) Each gene is then corrected with array gene centering (AGC).

8.1.2. Annotation

In the data annotation part the first step was to manually curate the annotation of each sample. For this purpose we defined 37 annotation fields to be filled for each sample. Naturally, the accuracy of annotation depends on the quality and accuracy of the annotation provided by the original authors. For the primary annotation of GeneSapiens we manually curated altogether 175975 annotation details, excluding 6 fields with automatic filling by the data

importing algorithms. On average each sample has 18 manually curated annotation details. Table 1 describes the most important biological annotation fields used in primary annotation.

Table 1 Description of 17 biologically most relevant annotation fields used to describe each sample. The data analysis and secondary layer of annotation rely on these listed fields.

Name of the field	Description of the field	Percentage of samples having the information
Exp type	Divides samples into healthy, malignant and other disease classes	100
Anatomy	Anatomical system from which the sample originates	100
Cell type	Cell type from which the sample originates	60.5
ICD10	ICD10 code describing the pathological status of the sample	79.9
ICDO	ICD-O code further specifying the type of malignancy	67.6
Tissue preparation	Definition of sample preparation type	96.6
Age	Age of a person from which the sample originates	43.7
Sex	Sex of a person from which the sample originates	58.6
Ethnical background	Ethnical background of a person from which the sample originates	18.7
T-stage	Tumor-stage of a sample (malignant)	19.7
N-stage	Nodal-stage of a sample (malignant)	24.4
M-stage	Metastasis-stage of a sample (malignant)	21.3
Grade	Grade of a sample (malignant)	28.2
Histology	Detailed histological information of the sample	33.3
Vital status	Information whether the patient was alive or not at the time of the sampling	28.4
Survival day	Days the patient survived	13.2
Comment	Further notes about the sample	43.5

A secondary layer of annotation was done by grouping samples having specified combinations of primary annotations into groups of samples. This secondary layer of annotation formed 1215 groups of samples usable for data analysis purposes. The secondary layer of annotation specifically allowed relatively easy way of analyzing samples on various levels of biological systems (e.g. breast cancer as one group and various histological subtypes of breast cancer as separate groups).

8.2. Validation of GeneSapiens

GeneSapiens data was validated with both mathematical and biological methods.

8.2.1. Mathematical validation

The issue of defining goodness of normalization as well as goodness of entire data integration procedure was one of the key questions to be solved. On the side of more mathematical validation we found out that multidimensional scaling (MDS) can provide a rather good overview as it is able to provide visually informative plot of the overall clustering tendency of the data. The original problem of array generation driven clustering of data is clearly visible from the MDS analysis as well as is the effect of the normalization. This is demonstrated in Figure 6 which shows a 2-dimensional representation of a Pearson correlation coefficient based distance matrix of 1489 healthy *in vivo* samples. The same set of samples has been analyzed after three different normalization steps (MAS5, EQ and AGC). It is clear that when this set of samples has been preprocessed with MAS5 the clusters of samples are defined largely by the array generations (Figure 6A). After EQ, the clusters are perhaps slightly more compact, but still defined by array generations (Figure 6B). However, after AGC correction we can see a dramatic mixing of array generations among the clusters suggesting that some other feature of the data defines the clusters (Figure 6C). When we examine the clusters of AGC corrected data in terms of anatomical origin of samples we can see that clusters are formed definitely more in terms of anatomical origin of the sample than the array generation of the sample (Figure 6D).

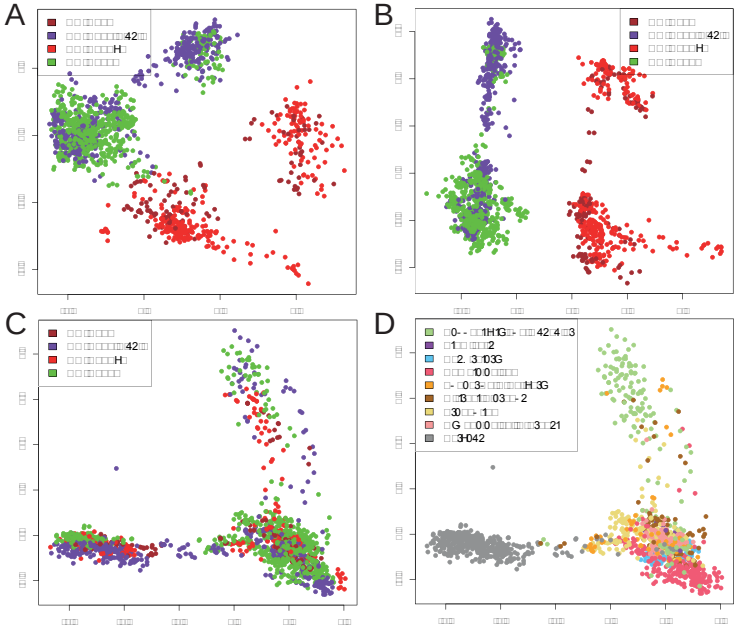

Figure 6 Two-dimensional multidimensional scaling of Pearson correlation coefficient based distance matrix between 1489 healthy *in vivo* samples. A) Samples after MAS5 preprocessing colored by the used array generation B) Samples after EQ normalization colored by the used array generation C) Samples after AGC correction colored by the used array generation D) Samples after AGC correction colored by the anatomical origin of the sample.

In publication I we performed a more exact quantification of the sample clustering before and after AGC correction by K-means clustering of the data. Subsequent rand index calculation [116] revealed that the index of sample segregation into the array generation based clusters decreased from 0.45 to 0.15 (where 1 indicates perfect segregation) after the correction. Conversely sample segregation based on the anatomical system of origin increased from 0.22 to 0.92.

In publication I we also studied correlations between technical replicates of the same samples analyzed by different Affymetrix array generations. Correlation coefficients between 14 human muscle biopsy samples analyzed on U95Av2 and U133A array types increased to >0.9 after AGC correction whereas before the correction coefficients were <0.75. Another study contained 123 leukemia samples analyzed on three array types (U95Av2, U133A and U133B) [123, 124]. We found out that mean value of correlation coefficients between the replicates was significantly higher (0.78) than mean value before the correction (0.5).

In publication II we further tested the performances of various normalizations with multiple measures of goodness. Altogether 6926 Affymetrix samples were preprocessed with MAS5 and then normalized by five different methods; standardization (Z), housekeeping gene based (HK), equalization transformation (Q), Weibull distribution based normalization (WBL), array generation based gene centering (AGC). Out of these we formed ten different combinations i) pure preprocessed MAS5.0 ii) Z iii) Q iv) WBL v) HK and all these combined with AGC (MAS5AGC, ZAGC, HKAGC, QAGC, WBLAGC). This collection of data contained more than 500 samples from five different Affymetrix array generations (HU6800, U95A, U95Av2, U133A, U133 Plus 2).

The goodness of normalization was defined in five different ways 1) correlation between technical replicates 2) correlation between randomly selected genes 3) classification of the samples based on the anatomical class 4) comparison of correlations between the samples computed based on the anatomical classes and array generations 5) stability of the housekeeping gene expression levels.

First measure of normalization goodness was done by calculating correlation coefficients between technical replicates by using same set of 123 leukemia samples [123, 124] as was used in the publication I. This method of data comparability has been used in several studies [56, 57, 71]. MAS5, Z and HK gave identical results since the methods are linearly invariant while WBL gave slightly better correlation coefficient. However, combining any of these methods with AGC resulted in significant increase in correlation coefficient value (Publication II, Figure 1). WBLAGC gave the best correlation coefficient but the difference to other AGC combined methods was not significant.

The second measure of normalization goodness was based on testing correlation coefficients of randomly selected gene pairs. The assumption here is that randomly selected genes should not be correlated with each other, thus the expected value for their correlation is zero. Technical variation from different array generations may cause systematic errors in the data, which

reflects as non-zero correlation between random gene pairs. We calculated correlation coefficients between 500 randomly chosen genes, each having values in all array generations, from data normalized with previously described ten methods. Results showed that ZAGC, HKAGC, QAGC and WBLAGC had values significantly closer to zero than other methods (Publication II Figure 2).

The third measure of goodness of normalization assumed that the expected value of correlation coefficient between samples from the same anatomical class is higher than the expected value of correlation coefficient of samples from different anatomical classes, even if the samples originate from same array generation or experiment series. For this purpose we calculated correlation between all 1464 samples and divided the results in two groups 1) correlations of samples from the same array generation but different anatomical class 2) correlations of samples from the same anatomical class but different array generations. Results showed that without AGC the array generation was stronger than anatomical class in defining the identity of the sample. However, when AGC was used correlation coefficients between the samples of same anatomical origin had significantly higher correlations than between samples from different anatomical origin and from the same array generation (Publication II Figure 3).

Stability of housekeeping genes was used as the fourth measure of normalization goodness. This is based on the assumption that there is a set of genes of whose products are needed for the basic metabolism of all cells and therefore the expression levels of these genes are relatively stable. We used 126 of these so-called housekeeping genes [24] to measure the goodness of normalization by assuming that the better the normalization the more stable the housekeeping gene expression is. For each of the genes we calculated Kullback-Leibler distance between the distribution of expression values in one array type and the distribution of expression values across all array types. After AGC, normalization distributions of housekeeping genes in the different array generations resemble more the combined distribution across all array generations, in other words, the expression of housekeeping genes is more stable after AGC (Publication II Figure 4).

The fifth measure of normalization goodness was performed by classifying samples to the anatomical systems of origin with nearest neighbor classifiers. For each anatomical system of origin (n=35) the mean expression profile of logarithmic expression values was calculated, with each anatomical system having at least 10 samples. This analysis was limited to 1464 healthy tissue samples present in the dataset, as the malignant tissue samples might have caused severe bias for classification according to anatomy. Each sample was then classified into most similar anatomical system with nearest neighbour algorithm. Normalization methods with AGC achieved substantially higher accuracies with average accuracy of 89% versus 76% without AGC (Publication II Table 1).

8.2.2. Biological validation

In publication I we used genes with known tissue specific expression profiles to validate the data integration procedure. Figure 7 summarizes the results for

seven genes with well-known tissues-specific expression. Troponin T type 2 (*TNNT2)* shows very clear heart specific expression, as expected for a clinically used cardiac biomarker, even though data originated from four different array generations and comprises only about 0.5% of the samples. High expression of *TNNT2* in rhabdomyosarcoma was also found out to correspond with observation of increased troponin T levels in serum of rhabdomyosarcoma patient [125]. Similarly the body wide expression profile of placental alkaline phosphatase (*ALPP)* shows the known expression in placenta [126] and also confirmed the known ectopic expression [127, 128] in various types of cancers. Myelin-associated glycoprotein (*MAG)*, a known neuronal marker [129] shows high expression in healthy central nervous system and to some extent in gliomas. Similarly Kallikrein- 3 (*KLK3)*, Glial fibrillary acidic protein (*GFAP)*, Insulin (*INS)* and L-lactate dehydrogenase C (*LDHC)* show corresponding known tissue specific expression in prostate [130], nervous system [131], pancreas and testis [132], respectively. Further biological validation was acquired by using GeneSapiens for studies III and IV as well as various collaborative research projects.

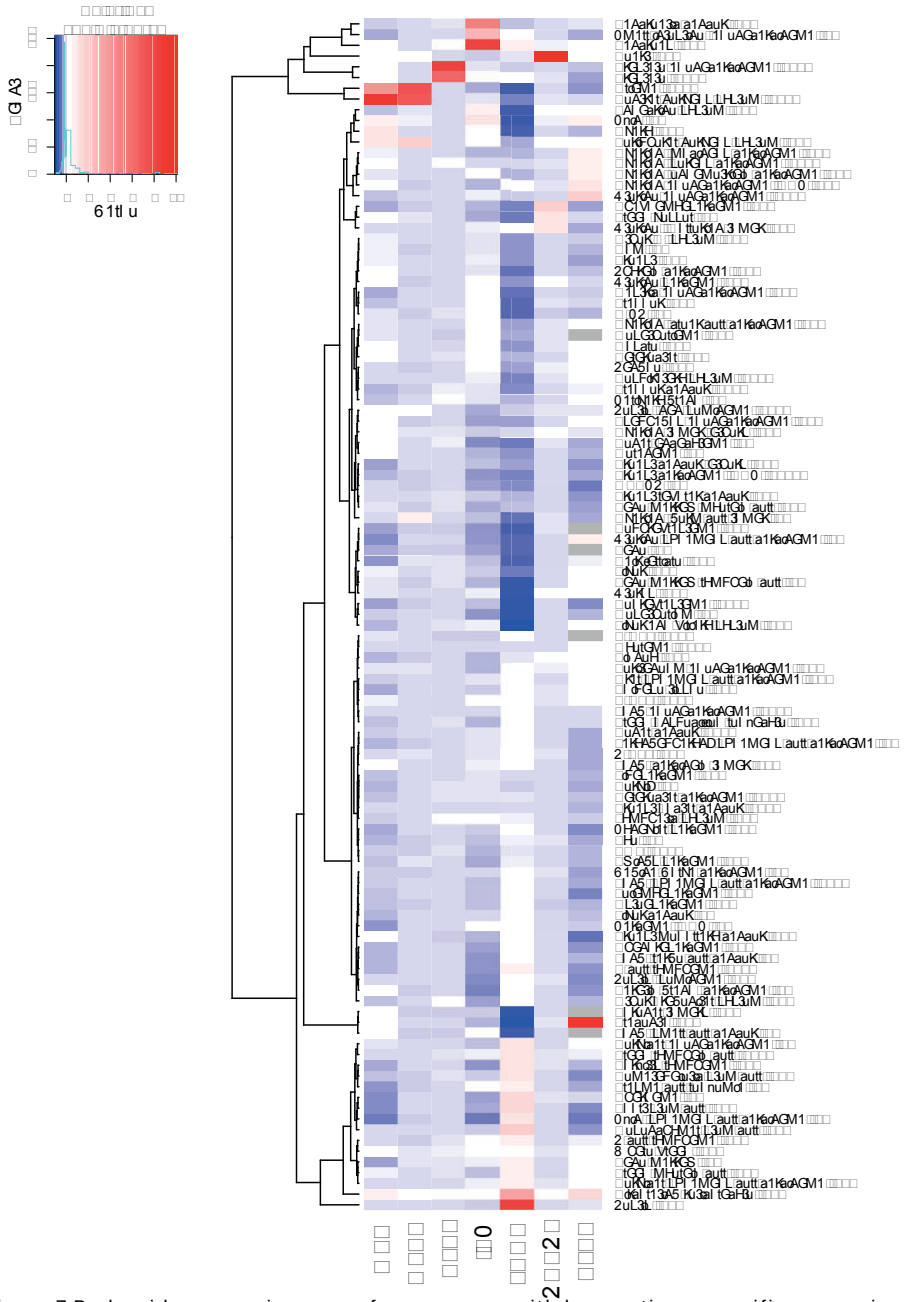

Figure 7 Body wide expression map of seven genes with known tissue specific expression. On the x-axis are genes (*MAG*, *GFAP*, *KLK3*, *INS*, *LDHC*, *TNNT2* and *ALPP*) whereas on the y-axis are 111 healthy and malignant tissues (number of samples per tissue is given in the parentheses). Data has been scaled genewise and both axes have been clustered hierarchically with Euclidean distance and Ward linkage. Color indicates the average expression of the gene in the tissue in question.

8.3. Application of GeneSapiens data

GeneSapiens data is essentially two dimensional, the expression values of 17330 genes across 9783 samples, with some covariate data associated to both dimensions. Therefore the application of GeneSapiens data is best to be understood as analyses of either the gene or sample dimension.

8.3.1. Gene dimension analyses

The first application of GeneSapiens data was to study how various individual genes are expressed across the whole human body. In publication I we showed the bodywide expression profile of preferentially expressed antigen in melanoma (*PRAME)*, a gene having high expression in healthy testis but also showing high ectopic expression in various human cancers (Publication I Figure 4). Even though the high expression in various human cancers was already known [133] the benefit of large-scale data integration is obvious as it comprehensively reveals all the tissues where the gene is expressed. One of the key challenges of genome wide gene expression measurements has been the difficulty of finding a proper reference for the interpretation of expression values. An ability to put expression level into the context of thousands of other properly annotated samples provides a robust reference for interpretation. Specifically the body wide dot plot allows visualization of thousands of data points in a biologically sensible manner.

In publication I we performed gene family analysis by creating a bodymap of expression values of genes listed in Sanger Center cancer gene census (Publication I Figure 5). This heatmap style of data analysis and visualization allows one to identify subgroups of genes or tissues having common expression profiles. In principle, the bodymap combines both the gene and sample dimension analyses of GeneSapiens data. As one might imagine the expression profiles of human cancer genes divided tissues into malignant (84.4% malignant tissues), healthy (82.1% healthy tissues) and to hematological (100% hematological tissues) clusters. Somewhat surprising was the observation that hematological cluster contained both healthy and malignant tissues without apparent division. The gene dimension revealed five distinct clusters of cancer genes. To interpret potential functional differences of gene clusters we calculated correlation coefficients between the cancer genes and known marker genes antigen KI-67 (*MKI67*), proliferating cell nuclear antigen (*PCNA*), cytokeratin-19 (*KRT19*) and leukocyte common antigen (*PTPRC*) across the entire database. Cancer genes correlating with proliferation markers, *MKI67* and *PCNA*, were the ones with unusually high expression in solid malignancies. *KRT19* correlated mainly with epithelial cancers. And as one might assume genes correlating with the hematological marker *PTPRC* were most often expressed in hematological malignancies.

The bodymap of human cancer genes pinpointed several genes with extremely high expression. One example of this was the Mast/stem cell growth factor receptor (*KIT)* with extremely high expression in gastrointestinal stromal tumors (GIST) (Publication I Figures 5 and 6). This is a prime example of how having a comprehensive expression profile allows one to identify the extreme high expression of a gene in a malignancy of which the healthy counterpart samples are hard to get. It suggests that the gene in question might have a fundamental role in the tissue of question and therefore might provide novel

therapeutic angles. High expression in this case most likely is not cancer specific, as it is likely that interstitial cells of Cajal, from which GIST is though to originate [134], also express *KIT* at high level, but the extreme expression suggests that these cells are highly dependent on *KIT*. GIST patients often have mutated *KIT* and Gleevec has been found to be effective drug targeting *KIT* in addition to its primary target, the *BCR-ABL* fusion gene. Similarly bodywide expression profiles of the ETS oncogene family (*FEV)* and the Myeloid/lymphoid or mixed-lineage leukemia *(MLLT11)* allowed us to identify high expression in various malignancies. The key conclusion is that many of these observations are only possible with the ability to interpret expression levels in the context of a bodywide collection.

In publication III we performed an even more extensive analysis of the bodywide expression profiles of 459 human kinases. We defined the concept of transcriptionally active genes by analyzing the expression levels of kinase genes across 1603 healthy samples representing 44 tissues. By using entropy based methodology we were able to define an expression level threshold for each kinase gene above which the gene can be assumed to be under active and positive regulation. This is based on the assumption that after integrating data from multiple tissues together we are able to define the background expression level shared by most tissues. After defining the transcriptional activity thresholds for all kinase genes we were able to binarize the expression levels of kinase genes over 99 tissues into transcriptionally active/inactive states (Publication III Figure 1A). Hierarchical clustering of the tissues revealed that the binary expression state of human kinome is informative enough to create an overall similar threefold separation to healthy, malignant and hematological tissues as seen in the bodymap of human cancer genes (Publication III Figure 1B). However, an additional mixed cluster was formed containing some healthy-malignant pairs not separable from each others based on their kinase profile. There were also distinct subclusters of neuronal and muscular tissues among the healthy ones whereas solid tumors had subclusters of non-epithelial and epithelial tumors, the latter further subdiving into adeno- and squamous types. Interestingly, the human kinome, which is strongly regulated at the protein level, is also under strong enough transcriptomic regulation to define tissue type only based on the profile of transcriptionally active kinase genes. Distinct clustering of various tissue types suggested that transcriptional activity levels provided biologically solid data.

We were also able to identify several groups of kinases having distinct transcriptional activity profiles across the tissues (Publication III Table 1). The most prominent one was named “proliferation” kinase genes as it was mainly active in solid tumors and immunological/hematological tissues. Kinases of that group were transcriptionally active in 88.7% of solid cancers and in 65.8% of immunological/hematological tissues. In the healthy and mixed tissue groups the percentages were 20.8% and 44.2%, respectively. Other identified groups of kinase genes included “hematological”, “neuronal, ”non-epithelial”, “epithelial” and “generally active”.

The same methodology of defining transcriptional activity can be applied across the entire human transcriptome to define which genes are assumed to be under active transcriptional regulation in each tissue. Figure 8 summarizes

the results of such an analysis performed for each gene in all healthy tissues. The figure itself provides only an overview of the transcriptomic activity of human genome across 44 healthy tissues.

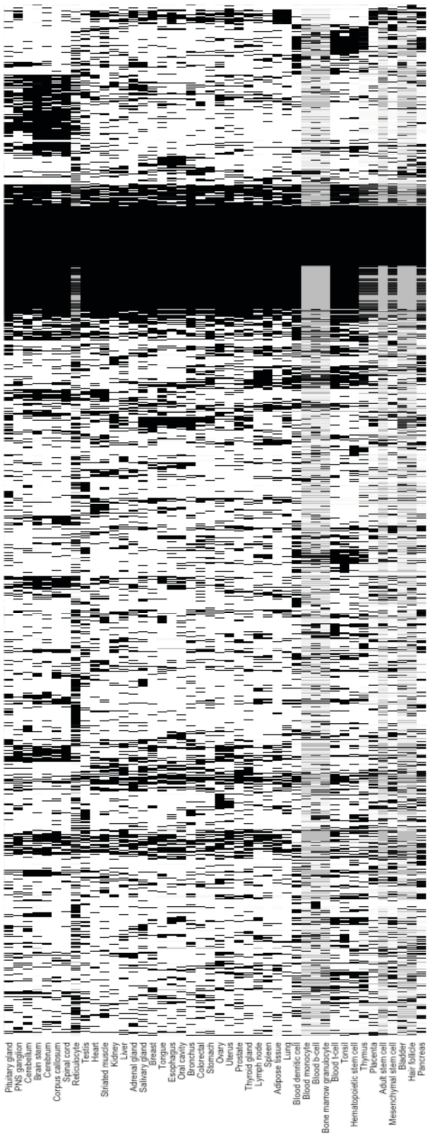

Figure 8 Transcriptional activity map of 17 330 genes across 44 healthy human tissues. Black indicates that a gene is transcriptionally active in the tissue in question. Grey indicates missing data for a gene in the tissue in question. Both axes have been clustered with binary distance metric and complete linkage method. Figure is provided only as an overview of the transcriptomic activity as the amount of data greatly surpasses the resolution of the printed image.

Further analysis of the transcriptional activity levels shows that the majority of genes are transcriptionally active in less than 50% of healthy tissues but there are also nearly 2000 genes (11.5% of all genes) transcriptionally active in virtually all tissues (Figure 9A). This number is slightly less than recently estimated by Zhu *et al.*[135], but it is largely a threshold issue. In publication III clustering of human tissues in terms of binary transriptional activity profile revealed distinct clusters. Similarly, usage of the binary transcriptomic activity profiles of all genes as a basis for clustering of 44 healthy tissues revealed an anatomically logical clustering of tissues (Figure 9B).

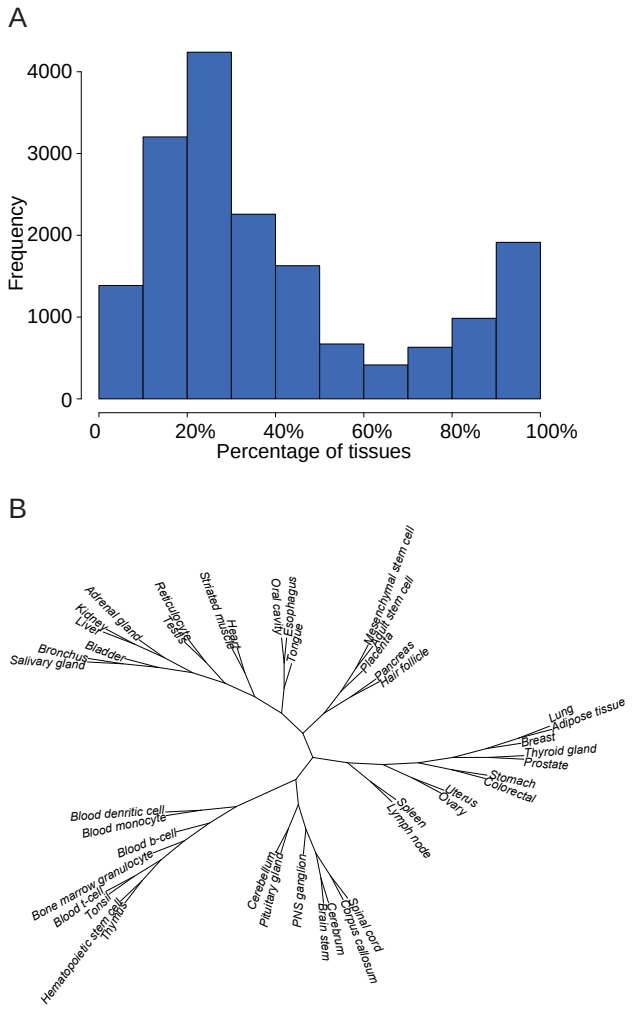

Figure 9 Characteristics of healthy transcriptome across 44 tissues. A) Number of genes transcriptionally active in healthy human tissues presented as percentage of all 44 tissues present in the study. For example there are approximately 2000 genes that are active in at least 90% of the tissues. B) Similarities between 44 healthy human tissues in terms of their binary profile of transcriptionally active genes. The tree has been build upon binary distances with Ward linkage method and with the Analyses of Phylogenetics and Evolution (ape) R library [136].

Transcriptional activity status can also be easily used to identify new biomarkers. An example of such an analysis is shown in Figure 10 consisting body wide expression profiles of the top 50 most healthy tissue specific genes. The group of genes has been selected by requiring that they are transcriptionally active in no more than three healthy tissues and further prioritized by highest median expression level above the transcriptional activity level threshold. Analyses like this easily reveal known tissue biomarkers like *KLK3*, Tissue kallikrein-2 *(KLK2)*, *TNNT2* and Troponin I *(TNNI1)* and numerous less known or novel biomarker genes.

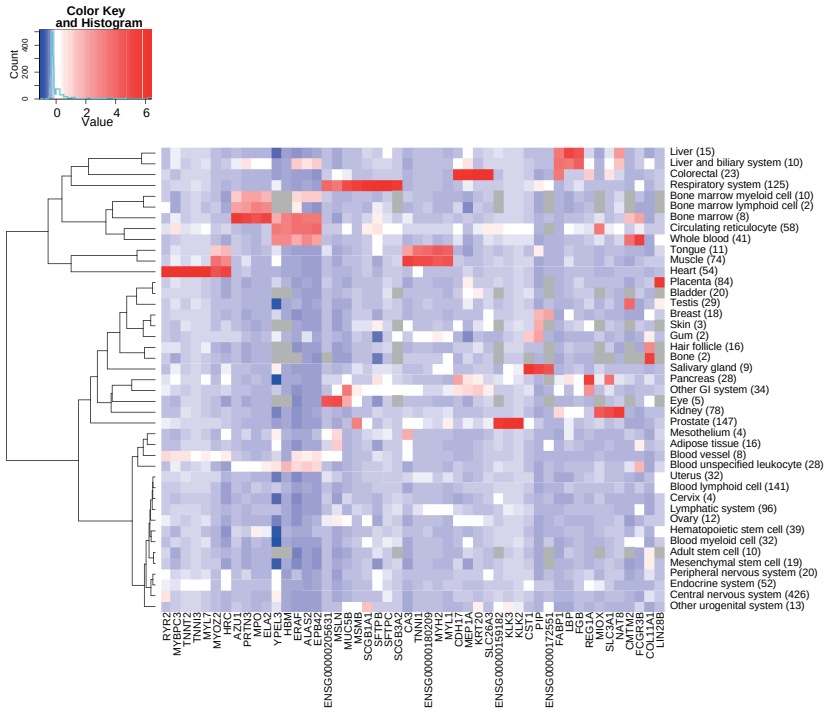

Figure 10 A bodymap of the top 50 most tissue specific genes. On the x-axis are genes whereas on the y-axis are 44 healthy tissues (number of samples per tissue is given in parentheses). Data has been scaled genewise and both axes have been clustered hierarchically with Euclidean distance and Ward linkage. Color indicates the average expression of the gene in the tissue in question.

In publication III we also studied the genomic co-expression environment of all kinase genes in order to find out with which biological processes they associate (Publication III Figure 2). Results showed that for most of the human kinase genes it is possible to find significant enrichments of biological processes, as defined by the Gene Ontology (GO-BP) [137], present in the genomic co-expression network around the kinase genes. Proliferation related kinase genes identified from transcriptional activity data were found to associate with DNA repair, cell cycle control, mitotic chromosome handling, chromatin handling and regulation of cell growth. This group of genes included the well-known mitotic kinase genes like Aurora kinase A *(AURKA)* [138], mitotic checkpoint serine/threonine-protein kinase (*BUB1)* [139], Polo-like kinase 1 (*PLK1)* [140], Dual specificity protein kinase (*TTK)* [141], cyclin-dependent kinase 1 (*CDC2)* [142], MAPKK-like protein kinase *(PBK)* [143], Mitotic checkpoint serine/threonine-protein kinase beta *(BUB1B)* [144], Polo-like kinase 4 *(PLK4)* [145], NimA-related protein kinase 2 *(NEK2)* [146], Serine/threonine-protein kinase Chk1 *(CHEK1)* [147], Aurora kinase B *(AURKB)* [148], Cell division protein kinase 2 *(CDK2)* [148], but also several novel ones like Microtubule-associated serine/threonine-protein kinase-like *(MASTL),* Maternal embryonic leucine zipper kinase *(MELK),* Dual specificity tyrosine-phosphorylation-regulated kinase 2 *(DYRK2),* DNA-dependent protein kinase catalytic subunit *(PRKDC)* which are not yet experimentally proven to be mitosis and/or cell cycle related.

8.3.2. Sample dimension analyses

In publication IV we explored a methodology allowing one to "align" an expression profile to a reference database of expression profiles. The specific aim of this capability is to enable interpretation of the "query" expression profile in the context of the existing expression profiles. This resulted in the development of the method for alignment of gene expression profiles (AGEP). The principles of the method are explained in Figure 11.

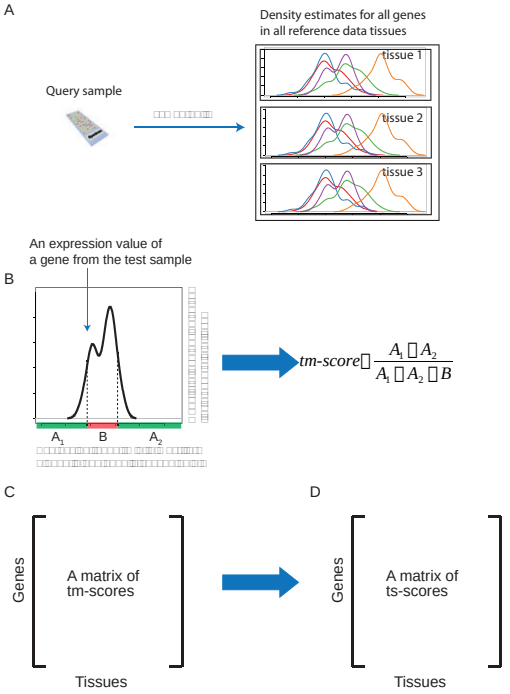

Figure 11 Schema of the AGEP method. A) An external test sample is compared to density distributions calculated from the reference data tissues. B) At the individual gene level the method calculates the range of the density estimate where the density is lower than the density value of the expression in the test sample. The result is called as the tissue match (tm-) score (0<tm-score<1). C) The calculation is repeated for all genes of the test sample against all reference tissues resulting in a matrix of tm-scores. D) The tm-scores of each gene against all reference tissues are further compared together to analyze uniqueness of the tm-score value resulting in a matrix of tissue specificity (ts-) scores (-1<ts-score<1).

Comparing an individual test sample against the reference database results in two matrices of gene and reference tissue specific scores (tm- and ts-scores), as seen in Figure 11C-D. In short, tissue match scores (tm-scores) describe how well the expression levels of genes in the test sample match the expected expression levels in the reference tissues. The tm-scores vary between 0 and 1, where 0 means "no-match" and 1 means "perfect match". Tissue specificity scores (ts-scores) describe whether the match was tissue specific or not. The ts-scores vary between -1 and 1, where -1 means "not matching tissue specific expression level" and 1 means "matching tissue specific expression level". In publication IV we showed multiple ways how these scores can be used to interpret the nature of the test sample expression profile. See material and methods and publication IV for further details.

As a validation of the accuracy of the method we tested its ability to identify the tissue of origin of the query samples (origin defined as the reference tissue with highest average ts-scores). This was done both as a leave-one-out cross-validation (LOOCV) analysis [149] for the entire reference database and for an external dataset of 195 samples from the Array Express study E-GEOD-7307. In LOOCV analysis all samples of the reference database were tested one by one as if they were external samples. It is essential for LOOCV that the sample

being tested is removed from the reference data before testing. Overall accuracy with LOOCV was 93.6% (with a range of 58.3-100% depending on tissue type) and with external dataset tissue of origin was correctly identified for 84.6% of samples and for 12.3% of samples a closely related tissue type was identified (actually this 12.3% of samples consists entirely central nervous system (CNS) samples where different anatomical parts of brain are mixed) (Publication IV Table 1 and Supplementary figure 2). Altogether AGEP was found to be at least as accurate in identifying the tissue of origin as the more simpler nearest-neighbor (NN) [150, 151] and the more complex support vector machine (SVM) [152-154] based algorithms (Publication IV Table 1).

One of the key features of AGEP is its ability to quantify the similarity of the query profile to the reference tissues at the level of genes. This allows one to study gradual changes in the transcriptomic program for example during the differentiation of cells. In this particular example seven samples, each representing a distinct state of myeloid cell differentiation, were compared against hematopoietic stem cell (HSC), granulocyte and monocyte reference tissue classes (Publication IV Figure 6). As one might assume, the hematopoietic stem cell sample had the largest amount of genes having HSC stem cell specific expression levels. In a somewhat more differentiated myeloblast sample HSC specific expression levels of genes disappear and the sample represents more “in between” type of profile than anything specific. In the monocyte differentiation line in the monoblast sample some genes gain monocyte specific expression levels and resembles monocytes somewhat more, whereas the monocyte sample, as presumed, has the most monocyte specific expression levels. Similarly when moving from myeloblast to granulocyte the sample gets most granulocyte specific expression levels. On the malignant side, interestingly, the leukemia stem cell sample resembles HSC, but an AML sample seems to resemble a non-stem cell like transcriptome with differentiation direction not similar to granulocyte or monocyte.

We also used AGEP to compare a series of samples from an experiment of the differentiation of mesenchymal stem cells (MSC) against the reference database (Publication IV Figure 4-5 and Supplementary figure 4). The aim of the experiment had been to study of adipogeneic differentiation and we were able to show that samples change their transcriptomic program towards adipose tissue during the differentiation process. AGEP also allowed us to identify key genes changing during the differentiation. Proteoglycan link protein *(HAPLN1)*, Stanniocalcin-2 *(STC2)*, Ajuba isoform 2 (*JUB)* and Dickkopf-related protein 1 *(DKK1)* had relatively MSC cell specific expression levels in all replicate samples at timepoint 0h. After seven days of differentiation Adiponectin precursor *(ADIPOQ)*, Lipid droplet-associated protein *(PLIN)*, Thyroid hormone-inducible hepatic protein *(THRSP)* and MOSC domain-containing protein 1 *(MOSC)* gained expression levels matching expression levels relatively unique for adipose tissue. All these are previously known to be adipose tissue related genes [155-158].

The change in the expression level of *ADIPOQ* is shown in more detail in the Figure 12. At the 0h timepoint all replicates had such an expression level of ADIPOQ that it clearly matched the expression level commonly observed in

many tissues. At the 7d timepoint all replicates had gained an adipose tissue specific expression level for the gene. Table 2 summarizes how these changes are reflected in the scores provided by the AGEP method when comparing sample replicate A against MSC and adipose tissue at 0h and 7d timepoints in terms of *ADIPOQ* gene. The key observation is that the tm-score for MSC drops from 0.99 to 0, reflecting the loss of an expression level matching MSC (and most of the other tissues). Similarly, the ts-score for adipose tissue increases from -0.2 to 0.9, reflecting the gain of adipose tissue specific expression level. The relatively high tm-score against adipose tissue at the 0h timepoint (0.78) points to some over sensitivity of the method to emphasize the meaning of individual data points of the reference data. The most likely cause for the high tm-score in this case is a single adipose tissue sample within the reference data with low expression for the gene in question.

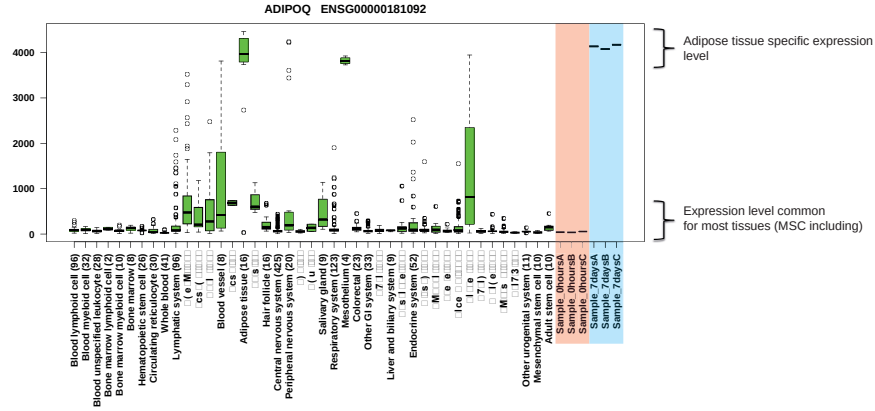

Figure 12 *ADIPOQ* expression levels in normal tissues, with 6 mesenchymal stem cell (MSC) samples included (on the right). There were three replicates of timepoint 0h samples, highlighted in light red, and three timepoint 7d samples, highlighted in light blue. At timepoint 0h each sample had *ADIPOQ* expression level typical for most of the tissues. On the contrary, after seven days of differentiation towards adipocytes, samples obtained adipose tissue specific expression levels.

Table 2 Tissue match and tissue specificity scores of ADIPOQ gene when comparing MSC replicate A against mesenchymal stem cells and adipose tissue at 0h and 7d timepoints.

	Timepoint 0h	Timepoint 7d
MSC tm-score	0.99	0
MSC ts-score	0.02	-0.18
Adipose tissue tm-score	0.78	0.95
Adipose tissue ts-score	-0.20	0.9

In publication IV we also showed how analyses at the individual gene level can be extended to gene sets to understand the changing transcriptomic program at the level of biological processes or pathways. For example, in terms of the geneset of adipose tissue differentiation and the geneset of lipid and fatty acid transport the 0h time point sample had expression levels mostly matching those typical for MSC. After seven days of differentiation, samples mostly resemble adipose tissue at the tissue level with both genesets having expression levels matching adipose tissue (Publication IV Figure 4).

9. Discussion

The need for data integration is as natural as the scientific understanding of complex phenomena is possible only with comprehensive and accurate data collections. The integration of large amounts of transcriptomic data has been the aim of many studies and many have also succeeded in it [17-20, 80-83, 85], gaining further understanding of biology of gene expression.

The results presented here comprise one of the most comprehensive and most integrated transcriptomic data sources constructed so far. In publication I we described how GeneSapiens database was constructed from 9873 *in vivo* samples. The major challenges in the transcriptomic data integration were the variance in data annotation nomenclature as well as the technical variation in how different array generations measure expression of genes. The enormous complexity of biological organisms ensures that no two samples are exactly alike, however there needs to be some sort of systematic way to describe what kind of sample has been hybridized on the microarray, so that at least semi-equivalent samples can be identified. Additionally, for biologically sensible datamining one needs the ability to identify samples belonging to various hierarchical levels of biological classification of anatomical systems, organs and tissues. These requirements render current computational annotation methods ineffective and inaccurate as there are not enoughly sophisticated algorithms to handle true complexity of the sample annotation. Thus, we used manual curation as the main method of annotation. Automation with computational tools was used as much as possible without compromising accuracy. Altogether this scheme of annotation has most likely produced the most accurate collection of transcriptomic data so far. However, one must understand that the collected and integrated data is essentially only as reliable and usable as its annotation in the original study is. We noticed large quality differences in the annotation details given by the original authors, also newer studies were found to have a significantly higher quality of annotation data available when compared to studies published several years ago. It is hard to say whether it is due to standards like MIAME [75] being slowly adapted by the scientific community or general increased awareness that expression data could be useful also in contexts other than the original experiment setup.

The major source of technical variation was different hybridization kinetics of varying probe sequences in different array generations, as pointed out by Hwang *et al.* [56], Elo *et al.* [71], Canales *et al.* [53], Nimgaonkar *et al.* [57], Mecham *et al.* [72], Autio *et al.* [58]. The breakthrough in our research was the discovery that after having enough expression data across a multitude of tissues from each array generation one could actually observe the entire biologically occurring spectrum of expression values for a gene. This leads to the possibility of identifying technical variation in the array generations' ability to measure the expression of a gene in the form of differences in the distributions of expression values of the gene between array generations. Apart from the solution presented by Hwang *et al.* and Elo *et al.* [56, 71] where data from multiple array generations are integrated by using only probes with common sequences, the AGC correction presented in the publication I remains the only known solution for the problem. The largest hazard of this type of data integration is that if the expression data collection

is not large enough then expression distributions used in the correction inadequately represent the complete biological variation of the expression of genes. In this kind of case some part of the biological variation would be corrected as if it were technical variation and therefore bias would occur in the data. Unfortunately, there is no easy way to know when this happens, but the larger the used data collection is, the less likely it is to happen. It will be a highly interesting avenue of research to find out how much data is actually needed to define the range of virtually all-possible biologically meaningful expression levels for each gene in each possible tissue.

Another potential hazard still present in the GeneSapiens data is variation caused by different sample preparation and hybridization protocols. Currently used normalization schemes are unable to compensate for that kind of variation if it causes non-linear changes between genes´ measured expression values. Another caveat is the type of sample taken from *in vivo* tissues; a bulk sample invariably contains multiple cell types whereas microdissected samples contain preferentially only single cell type. The former sample represents congregate expression levels over all the cell types whereas a microdissected sample contains more purely the expression levels of a single cell type, therefore microdissection can give much higher expression levels for genes specific for the single cell type when compared to bulk samples where other cell types average out the expression levels. This returns back to the accuracy of sample annotation, neither the bulk or microdissected samples should be assumed to be wrong in any particular way but should not be treated equally in terms of their representation of the tissue in question. Fortunately, large amount of data can give some level of quality control even against problems like this, since if a single study behaves completely differently from others it should be subjected to further scrutiny.

Understanding the expression levels among the large collection of data highlights the key advantage of integrating transcriptomic data. By comparing expression values to a large collection one is able to put those into context and understand what is really “high” and “low” from the otherwise relatively enigmatic values. The best example of this is the identification of biomarkers as shown in the publication I. *KLK3*, *MAG* and *TNNT2* could not be declared as tissue biomarkers with any reasonable level of specificity without knowledge of how the expression in the tissue in question relates to other tissues. With biomarkers, tissue or pathology specific expression is the key question, but when studying the expression levels of genes targeted by existing drugs highly interesting question is whether there is some other pathological condition with a similar expression level as the one already treated with the drug. In other words, if a gene is an optimal drug target in some pathological condition due to the high expression level, then one can ask is there some other disease with similarly high expression level that could also be treated with the same drug. The bodywide expression profile of *KIT* is an example of this kind of drug repositioning.

KIT is an interesting example also for the concept of transcriptional activity. As *KIT* is extremely highly expressed in GIST one can almost certainly assume that it is somehow active in that tissue. It does not allow for any assumption of the function at the protein level, but by comparing it to other tissues we can

assume that for some reason the gene's expression is actively and positively upregulated in that tissue. This kind of thinking lead us to publication III, where we found that we can define the likely expression level above which the genes are under active and positive transcription regulation. In that study we defined which human kinase genes are transcriptionally active over 99 healthy and malignant human tissues. Furthermore, we found out that kinase genes are under so effective a transcriptional regulation that the major types of human tissues can be identified solely based on the binarized profiles of kinase gene transcriptional activities. As kinases are known to play an essential role in the signal transduction of cells at the protein level this effective level of transcriptional regulation was one of the key new findings of the publication III.

The clustering of tissues based on transcriptionally active kinase genes revealed major clusters of solid healthy, hematological, solid malignancies and mixed clusters. Apart from the last-mentioned cluster, this is more or less similar to the division seen in the clustering made by using known cancer genes. Naturally the gene groups are partially overlapping but it is interesting to note that binarized expression profiles of kinase genes actually result in a more detailed clustering of cancers; even adeno and squamous types of cancers can be separated based on their kinase profiles. Our studies do not reveal whether this is mainly because kinases are important players in signal transduction and in malignant transformation or due to the noise reduction resulting from the binarization of the expression profiles. Additional analyses of clustering human tissues based on transcriptional activity levels of all genes support the fact that method provides accurate data across the entire genome. Therefore transcriptional activity levels could be used as foundation for further data mining.

As the definition of transcriptional activity does not allow us to assume anything else except that the gene most likely is under active and positive transcription regulation we set out to also study the functional associations of the genes' expression. This was done by forming a genomic coexpression network and subsequently finding which biological processes (GO-BP classes) were enriched in the vicinity of each kinase gene. These kinds of functional associations reveal if the expression of a kinase gene is correlated with the expression of group of genes all of which participate in the same biological processes and therefore, we could assume the kinase gene to be related into that process as well. Our results indicate that most kinase genes have rather clear associations to biological processes. The binarized expression profiles allowed us to make a rather straightforward comparative analysis of the kinase expression activity between healthy and malignant counterpart tissues by simply defining kinase genes whose transcriptional activity is lost or gained in malignant transformation. Functional associations allowed further understanding of what kind of biological processes are related into these changes.

However, the relation between gene expression, actual function and activity of proteins is not always entirely linear. For deeper understanding of the human kinome there needs to be both protein level and activity measurements as well as various functional screens to uncover biological functions of kinases.

Comprehensive analysis of protein levels and activities across tissues on the scale similar to genome wide expression analyses is technologically much harder to perform. On the technological side newly developed protein lysate arrays, as described by Leivonen *et al*. [91], are providing more high-throughput analysis of protein levels. However, large-scale kinase analyses with lysate arrays have yet to be performed. Other studies using more conventional immunohistochemistry (IHC) have been used to gather protein level data for increasing gene and tissue content. The most prominent of these studies is the human protein atlas described by Uhlen *et al.* [90]. They have profiled over 10 000 genes in terms of their protein levels over 46 normal human tissues and 20 cancer types. And as the protein levels do not correlate directly with the activity levels of proteins there is a need to separate protein activity screens. Various kinase activity profiling techniques are expertly reviewed by Johnson *et al*. [100].

On the side of functional screens Varjosalo *et al.* [99] cloned over 90% of full-length protein kinase cDNAs. They also constructed corresponding kinase activity-deficient mutant cell lines that can be used further study the effect of the kinase on the signalling network of the cells. Kinases have also been subject to protein sequence level analysis by Manning *et al.* [97] revealing highly interesting hierarchy of structural similarities among kinases.

Unforeseen understanding to the biology of human kinome might be achieved if data from i) DNA and protein sequence level analyses ii) transcriptomic analyses iii) protein level measurements iv) protein activity measurements v) results of functional screen could be fully integrated together.

In publication IV we explored the expression value distributions of genes beyond the simple binary concept as we developed a method allowing the comparison of single sample in the context of large reference database. Essentially we wanted to enable an analysis option analogous to BLAST (used in the analysis of nucleotide sequences). The need for these kinds of methods is great as an individual expression profile is actually surprisingly difficult to interpret. Theoretically, existing collections of expression profiles in public databases should provide ample reference material, but there are not too many methods allowing this kind of comparisons and interpretations. Methods unveiled in publication IV show how one can form expression value estimates for each gene in each tissue and then use this information as the basis for comparing an individual expression profile to a large reference database. The AGEP method allows quantification of the similarity between a single query sample and the sample groups (e.g. tissues) in the reference database. This similarity in the transcriptomic program can then be explored at the level of genesets and genes. Overall, AGEP achieved at least as good a tissue classification accuracy as the most commonly used advanced classifier methods like support vector machine (SVM) or algorithmically simpler ones like nearest-neighbour (NN). It is not sensitive to missing values and in an elegant way can handle the heterogeneity of the samples forming a group (e.g. tissue), an issue that is always present when complex biological samples are annotated and grouped together. Drawback of the AGEP method is somewhat substantial computational requirements.

AGEP can be categorized as a search & retrieval type of a tool, able to relatively quickly to search through a database of gene expression profiles and retrieve profiles most similar to the query profile. These kinds of methods have just begun to emerge, thus there are not too many reasonably comparable to AGEP. Perhaps one of the most similar method is the gene expression barcode published by Zilliox *et al*. [107]. The methodology generates a barcode of active/deactive expression of genes for each tissue. Thus the barcode in its simplicity provides merely binary information whether the gene is expressed or not in the tissue in question. A query sample can then be compared against barcodes of the reference tissues and the sample can be classified with rather high accuracy (over 90% in several independent datasets) as reported by Zilliox *et al.*. This definition of active/deactive expression states is broadly similar approach than the one we used in publication III. It is interesting to note that both publications emphasize the same fact; binary gene expression profiles are highly robust and informative.

An earlier related method has been reported by Parmigiani *et al*. [106], this statistical approach is broadly similar to the barcode method. The main difference between the two is that Parmigiani *et al.* considered genes in ternary mode (downregulated, normal and upregulated) when building a profile (analoguous to the barcode). These profiles can then be used to identify similarly behaving samples whether that similarity be a tissue classification or some other interesting property being studied.

Another related study has been reported by Caldas *et al.* [109]. They describe another search & retrieval type of a method, probabilistic retrieval and visualization of biologically relevant microarray experiments, which is able to search for related experiments from a large collection of gene expression studies. However, direct comparison with the AGEP method is somewhat challenging. AGEP has been designed to compare a single sample against a reference database without any *a priori* assumptions about genes. The method described by Caldas *et al*. compare entire experiment to other experiments in the reference database and the measure of similarity is based on geneset enrichment instead of expression of individual genes. The method also uses median as a summary value of the expression across the samples, thereby not addressing the multimodality issue that AGEP was designed to take into account.

One of the most interesting features of AGEP is that in the alignment of a query profile to the reference database one does not only get a similarity value for each reference sample class, but also a pair of values for each gene against each reference sample class. One of the values describes how well the expression of a gene matched the expected expression of the gene in a reference class in question and the other how unique this specific expression level of the reference sample class is. As the similarity to reference sample class is based on these values it is possible to understand how the similarity between query and reference class forms from the level of genes. Combining this with various paradigms of gene ontology or pathway style of gene annotation it is possible to uncover which biological processes component genes are expressed at typical level for which reference classes. For example, in publication IV we showed how differentiating MSC samples at the 0h

timepoint resemble the MSC reference class, but at the 7d timepoint they resemble the adipose tissue reference class. This change in transcriptomic program was further studied at the level of gene groups like one related to adipose tissue differentiation. At the 0h time point the genes of the group were expressed at level typical to MSC cells but at 7d timepoint they were expressed level typical to adipose tissue. Similarly, the change in the transcriptomic program was pointed at the level of individual genes as the genes with known role in adipose tissue gained adipose tissue specific expression level during the differentiation.

One fundamental difference of AGEP to many other expression datamining approaches is that it does not emphasize overexpressed genes more than down expressed genes, but on the contrary, it forms a surprisingly robust estimate of the observed expression levels for each gene. Additionally, it does not make any *a priori* assumption of the informative genes. In other words, the presented way of understanding and using expression data treats each gene as an individual entity. Each gene has certain potential distribution of expression available to it in the collection of samples grouped together under the assumption that they are biologically at least semi-equivalent. If the grouped samples are heterogenous in terms of expression of certain gene (e.g. histologically defined breast ductal cancer and *ERBB2* oncogene) then the distribution of potential expression values indeed is bimodal, as the gene in question can exist in two distinct expression states in the given population of samples. Given the heterogeneity of biological samples and the large number of genes it follows that practically all groups of samples will have some number of genes with bi- or multimodal expression distributions.

The bi- or multimodality of gene expression levels leads to an interesting hypothesis, slightly explored in the publication III: Do some or all genes have discrete expression states more often than continuous distribution of expression activity? In other words, are the usually observed continuous expression value distributions there just because of noise in our sampling and measuring systems, or could the cells' transcription regulation machinery actually be able to maintain more rigid control of the transcription? In the publication III we showed that the binarized expression states of human kinase genes seem to contain a great deal of information, and are able to reproduce large parts of the known and expected transcriptional behaviour of the genes. Thus it seems plausible to assume that at least for human kinase genes there are clearly two discrete states of expression. Results presented in Figure 8 and Figure 9 give supporting evidence that this could be true across the entire genome. Similarly, the study of Zilliox *et al.* [107] support the observation that it is possible to identify discrete expression states for all genes and the resulting data is accurate enough to perform tissue classification. In general, thresholding microarray data has been found to be a successful method as described by Shmulevich *et al.* and Pal *et al.* [159, 160]. Yet practically none of the existing studies touch deeply into the issue that is there more than two discrete states of expression.

In publication IV we noticed that approximately 16% of genes present in the study had bi- or multimodal expression distribution. The method used to estimate this number is completely different from the one used publication III

but it raises an interesting question. Does this 16% perhaps reflect a transcriptionally inconsistent sample grouping or could it be due to the cell's capability of maintaining multiple discrete expression states for genes? Naturally, this falls back to the annotation and logic of grouping samples together as representative of a certain specific tissue. From the binarization of gene expression status we learned that large-scale expression data supports two discrete states of expression. It remains an open question how many discrete states there are and what is the biological relevance of these states. Is it possible to build model of human transcriptome with multiple distinct 'digital' expression states for all genes across all tissues? Currently there are no studies properly addressing this issue, at least in the context of entire human transcriptome in hundreds of tissue types, but for certain such studies will be performed.

The ability to find expression based biomarkers, drug repositioning, identification of patient subgroups, exploration of expression activity of genes across wide spectrum of healthy and pathological tissues, functional associations through co-expression environment, interpreting individual expression profiles and construction of comprehensive models of human transcriptome have all in common the fact that they require large amounts of unified and integrated expression data. Data needs to be acquired and compiled together from multiple sources. In its entirety, the human transcriptome across healthy and pathological tissues is far too massive to be analysed in any single experiment.

10. Conclusions and future prospects

GeneSapiens contains manually curated annotation for 9873 *in vivo* samples with as complete clinical information as available, thus comprising the largest fully integrated gene expression database. It has been constructed by using solely Affymetrix microarrays, which are generally known to be very reliable and robust [7]. As pointed out by the MAQC study [5, 6] Affymetrix arrays have the best interlaboratory reproducibility. Additionally, GeneSapiens uses custom normalization designed to lower technical noise from various Affymetrix chip generations as well as uses state-of-the-art probe mapping [70] also pointed out by the MAQC study to be an important factor in reliability. A GeneSapiens type of resources can be used in understanding otherwise difficult to interpret expression values in both gene and sample dimensions. In the gene dimension one can identify tissues and/or pathologies where the gene is actively expressed or which genes are its co-expression partners. This is useful for multiple applications like basic research of gene functions, biomarker identification and drug repositioning as it provides a comprehensive "map" of gene expression activity across major portion of human tissues. On the sample dimension GeneSapiens data allows unforeseen possibilities to understand and interpret unknown or novel expression profiles. This line of thought has interesting consequences in the field of personalized oncology, the comparison of an expression profile from a patient's tumor against both healthy and pathological references could allow interpretation and understanding the expression level anomalies present in the tumor.

Methods like AGEP might provide a robust way to characterize all the expression anomalies of a patient's tumor in comparison to healthy reference. Furthermore, the robust ability to effectively compare a patient to a reference collection of thousands of patient profiles can situate the patient into the correct disease subtype and pinpoint expression anomalies that are personal to the patient in question. In theory these kinds of abilities could be used to prioritize treatment decisions on a truly personal level.

In overall, transcriptome research and meta-analysis should go towards more comprehensive modeling of the entire transcriptome. Even though GeneSapiens represents the most advanced fully integrated human transcriptomics database and can therefore push the research forward, the ramifications of deeper understanding of the structure and function of the transcriptome, as shown by Frith *et al.* [29] and Gingeras *et al.* [39] as well as the new possibilities opened by next-generation sequencing technologies, are already shaking the foundations functional genomics. At best, a gene and its measured expression level is only a suggestive model of the expression of the biologically meaningful and functional parts of the human DNA.

11. Acknowledgements

It is impossible to express my gratitude to all people contributing to this study in a way or another but this is my humble attempt. As this research spread over seven years it also spread over to many laboratories, universities and institutes. In terms of early steps of my Ph.D. research project, I would like to thank Biomedicum Biochip Center (as it was known at 2003) and newly founded Medical Biotechnology department of VTT. I want to thank also University of Turku and Technical University of Tampere for all support and infrastructure.

Later steps of the research were performed in the Genome Scale Biology research program in Biomedicum and in the Institute for Molecular Medicine Finland (FIMM). To FIMM I own sincere gratitude, it is great institute with even greater people. I want thank FIMM administration for taking good care of the researchers. Especially I would like to thank Reetta and Susanna. Altogether, I am very grateful to University of Helsinki for being positive part of my life and my career over 13 years. I am also grateful for the financial support received from Academy of Finland, Sigrid Juselius Foundation, Turku TE-center, Cancer Organizations of Finland and Helsinki University funds.

My most sincere gratitude goes to all people contributing this study, especially:

My supervisor Olli Kallioniemi, it has been a great privilege to work with him all these years. His almost supernatural ability to always invent new ways of thinking as well as novel approaches to problems has always and even still surprises me. He has also taken absolutely perfect care of providing the environment and financial support for the research. When working with him it is guaranteed that research is always reflected and linked to the state-of-the-art science that he seems almost always been aware of.

The official reviewers, Mauno Vihinen and Päivi Onkamo for their valuable comments about my thesis.

My coauthors, Reija Autio, Kristiina Iljin, Elmar Bucher, Henri Sara, Tommi Pisto, Matti Saarela, Rolf Skotheim, Mari Björkman, John-Patrick Mbindi, Saija Haapa-Paananen, Paula Vainio, Jaakko Astola, Matthias Nees and Sampsa Hautaniemi. It has been nice experience to work with you. Reija Autio is especially thanked for years of collaboration in challenging mathematical issues. Without Reija I wouldn't be here with this study completed. My heartful thanks, Reija. Elmar Bucher for his patience and hard work for collecting much of the used data. Henri Sara for his phenomenal ability to program and manage. Tommi Pisto for providing solution for one of the hardest challenges; how to share over 100 million datapoints with the rest of the world. Matti Saarela, whose mastery over mathematics and programming was essential for the research. Kristiina Iljin, Mari Björkman, Saija Haapa-Paananen, Paula Vainio, Rolf Skotheim, Kimmo Jaakkola and Kristine Kleivi for their essential work with data annotation. Matthias Nees for his thoughtful comments and discussions about the entire project. Sampsa Hautaniemi for

numerous enlightening discussions about bioinformatics and life in general. Outi Monni for providing support for the early steps of this research and further on always being friendly colleague and advisor. To all colleagues at FIMM for providing pleasant place to work.

Henrik Edgren for being a close colleague for all these years, it has been privilege to be able to discuss all kinds of things, science or not, with him. He somehow seems always to be able to relate ideas to known science and put things into the perspective.

Maija Wolf for contributing positively for everything. During these years she has been taking care and organized some many things that I often think how could anybody survive without Maija around. It has been very nice to work with her all these years. Especially I would like to express my gratitude to her in helping with the thesis writing.

Kalle Ojala for being my friend and part of my studies, my hobbies and my research since like an eternity ago. I am certain that I wouldn't be here without his contribution. Our deep and endless discussion and argumentation practically about every thought, idea and experience has taught me more than I can possibly realize. And he is damn good at putting some of my more abstract ideas into a formal mathematical form.

I want also to thank my friends Jack Leo, Simo Siiriä, Mikael Agopov and Mika Lindfors for participating in my hobbies and free time activities. Jack especially, as year more senior genetics student, has provided plenty of examples and helpful comments to follow in many aspects of studies and career in general.

My deepest gratitude goes to my loved one, Lotta, for making my life meaningful. She has endured all my up-and-down days of research, my sometimes somewhat long days and nights of work and my sometimes absent minded presence. Lotta makes my world so much better place to be.

Erityisesti tahdon kiittää myös vanhempiani kaikesta tuesta koulu- ja opiskeluajaltani. Kiitokset kuuluvat myös kummitädilleni. On erittäin todennäköistä, että vanhempieni jatkuva kannustus on mahdollistanut tämän nimenomaisen kirjan synnyn.

Helsinki, January 2011
Sami Kilpinen

12. References

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., et al., *Initial sequencing and analysis of the human genome.* Nature, 409(6822): p. 860-921. 2001.
2. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., et al., *The sequence of the human genome.* Science, 291(5507): p. 1304-51. 2001.
3. Istrail, S., Sutton, G.G., Florea, L., Halpern, A.L., Mobarry, C.M., et al., *Whole-genome shotgun assembly and comparison of human genome assemblies.* Proc Natl Acad Sci U S A, 101(7): p. 1916-21. 2004.
4. Schulze, A. and Downward, J., *Navigating gene expression using microarrays--a technology review.* Nat Cell Biol, 3(8): p. E190-5. 2001.
5. Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., et al., *The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.* Nat Biotechnol, 24(9): p. 1151-61. 2006.
6. Shi, L., Tong, W., Fang, H., Scherf, U., Han, J., et al., *Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential.* BMC Bioinformatics, 6 Suppl 2: p. S12. 2005.
7. Dalma-Weiszhausz, D.D., Warrington, J., Tanimoto, E.Y., and Miyada, C.G., *The affymetrix GeneChip platform: an overview.* Methods Enzymol, 410: p. 3-28. 2006.
8. Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 19(2): p. 185-93. 2003.
9. Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res, 31(4): p. e15. 2003.
10. Schadt, E.E., Li, C., Ellis, B., and Wong, W.H., *Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.* J Cell Biochem Suppl, Suppl 37: p. 120-5. 2001.
11. Workman, C., Jensen, L.J., Jarmer, H., Berka, R., Gautier, L., et al., *A new non-linear normalization method for reducing variability in DNA microarray experiments.* Genome Biol, 3(9): p. research0048. 2002.
12. Faller, D., Voss, H.U., Timmer, J., and Hobohm, U., *Normalization of DNA-microarray data by nonlinear correlation maximization.* J Comput Biol, 10(5): p. 751-62. 2003.
13. Rocca-Serra, P., Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., et al., *ArrayExpress: a public database of gene expression data at EBI.* C R Biol, 326(10-11): p. 1075-8. 2003.
14. Edgar, R., Domrachev, M., and Lash, A.E., *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res, 30(1): p. 207-10. 2002.
15. Hubble, J., Demeter, J., Jin, H., Mao, M., Nitzberg, M., et al., *Implementation of GenePattern within the Stanford Microarray Database.* Nucleic Acids Res, 37(Database issue): p. D898-901. 2009.
16. Day, A., Carlson, M.R., Dong, J., O'Connor B, D., and Nelson, S.F., *Celsius: a community resource for Affymetrix microarray data.* Genome Biol, 8(6): p. R112. 2007.
17. Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., and Pavlidis, P., *Coexpression analysis of human genes across many microarray data sets.* Genome Res, 14(6): p. 1085-94. 2004.
18. Rhodes, D.R., Kalyana-Sundaram, S., Mahavisno, V., Barrette, T.R., Ghosh, D., et al., *Mining for regulatory programs in the cancer transcriptome.* Nat Genet, 37(6): p. 579-83. 2005.
19. Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., et al., *ONCOMINE: a cancer microarray database and integrated data-mining platform.* Neoplasia (New York), 6(1): p. 1-6. 2004.
20. Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., et al., *Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.* Proc Natl Acad Sci U S A, 101(25): p. 9309-14. 2004.
21. Velculescu, V.E., Madden, S.L., Zhang, L., Lash, A.E., Yu, J., et al., *Analysis of human transcriptomes.* Nat Genet, 23(4): p. 387-8. 1999.
22. Warrington, J.A., Nair, A., Mahadevappa, M., and Tsyganskaya, M., *Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes.* Physiol Genomics, 2(3): p. 143-7. 2000.
23. Hsiao, L.L., Dangond, F., Yoshida, T., Hong, R., Jensen, R.V., et al., *A compendium of gene expression in normal human tissues.* Physiol Genomics, 7(2): p. 97-104. 2001.
24. Eisenberg, E. and Levanon, E.Y., *Human housekeeping genes are compact.* Trends Genet, 19(7): p. 362-5. 2003.
25. Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., et al., *Large-scale analysis of the human and mouse transcriptomes.* Proc Natl Acad Sci U S A, 99(7): p. 4465-70. 2002.
26. Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.* Proc Natl Acad Sci U S A, 101(16): p. 6062-7. 2004.
27. Shyamsundar, R., Kim, Y.H., Higgins, J.P., Montgomery, K., Jorden, M., et al., *A DNA microarray survey of gene expression in normal human tissues.* Genome Biol, 6(3): p. R22. 2005.
28. Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., et al., *Large-scale transcriptional activity in chromosomes 21 and 22.* Science, 296(5569): p. 916-9. 2002.
29. Frith, M.C., Pheasant, M., and Mattick, J.S., *The amazing complexity of the human transcriptome.* Eur J Hum Genet, 13(8): p. 894-7. 2005.
30. Lee, R.C., Feinbaum, R.L., and Ambros, V., *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 75(5): p. 843-54. 1993.
31. Ambros, V., *microRNAs: tiny regulators with great potential.* Cell, 107(7): p. 823-6. 2001.
32. Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T., *Identification of novel genes coding for small expressed RNAs.* Science, 294(5543): p. 853-8. 2001.
33. Mattick, J.S. and Makunin, I.V., *Small regulatory RNAs in mammals.* Hum Mol Genet, 14 Spec No 1: p. R121-32. 2005.
34. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 116(2): p. 281-97. 2004.

35. Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J., *miRBase: tools for microRNA genomics.* Nucleic Acids Res, 36(Database issue): p. D154-8. 2008.
36. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Res, 34(Database issue): p. D140-4. 2006.
37. Griffiths-Jones, S., *The microRNA Registry.* Nucleic Acids Res, 32(Database issue): p. D109-11. 2004.
38. Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., et al., *A uniform system for microRNA annotation.* RNA, 9(3): p. 277-9. 2003.
39. Gingeras, T.R., *Origin of phenotypes: genes and transcripts.* Genome Res, 17(6): p. 682-90. 2007.
40. Kapranov, P., *Studying chromosome-wide transcriptional networks: new insights into disease?* Genome Med, 1(5): p. 50. 2009.
41. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 5(7): p. 621-8. 2008.
42. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 11(1): p. 31-46.
43. Metzker, M.L., *Emerging technologies in DNA sequencing.* Genome Res, 15(12): p. 1767-76. 2005.
44. Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., et al., *A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.* Science, 321(5891): p. 956-60. 2008.
45. Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nat Genet, 40(12): p. 1413-5. 2008.
46. Holloway, A.J., van Laar, R.K., Tothill, R.W., and Bowtell, D.D., *Options available--from start to finish--for obtaining data from DNA microarrays II.* Nat Genet, 32 Suppl: p. 481-9. 2002.
47. Blanchard, A., Kaiser, R., and Hood, L., *High-density olinucleotide arrays.* Biosens Bioelectron, 11: p. 687-690. 1996.
48. Southern, E., Mir, K., and Shchepinov, M., *Molecular interactions on microarrays.* Nat Genet, 21(1 Suppl): p. 5-9. 1999.
49. Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P., et al., *Light-generated oligonucleotide arrays for rapid DNA sequence analysis.* Proc Natl Acad Sci U S A, 91(11): p. 5022-6. 1994.
50. Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., et al., *Light-directed, spatially addressable parallel chemical synthesis.* Science, 251(4995): p. 767-73. 1991.
51. Fan, J.B., Gunderson, K.L., Bibikova, M., Yeakley, J.M., Chen, J., et al., *Illumina universal bead arrays.* Methods Enzymol, 410: p. 57-73. 2006.
52. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhart, D.J., *High density synthetic oligonucleotide arrays.* Nat Genet, 21(1 Suppl): p. 20-4. 1999.
53. Canales, R.D., Luo, Y., Willey, J.C., Austermiller, B., Barbacioru, C.C., et al., *Evaluation of DNA microarray results with quantitative gene expression platforms.* Nat Biotechnol, 24(9): p. 1115-22. 2006.
54. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y., *RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.* Genome Res, 18(9): p. 1509-17. 2008.
55. Draghici, S., Khatri, P., Eklund, A.C., and Szallasi, Z., *Reliability and reproducibility issues in DNA microarray measurements.* Trends Genet, 22(2): p. 101-9. 2006.
56. Hwang, K.B., Kong, S.W., Greenberg, S.A., and Park, P.J., *Combining gene expression data from different generations of oligonucleotide arrays.* BMC Bioinformatics, 5: p. 159. 2004.
57. Nimgaonkar, A., Sanoudou, D., Butte, A.J., Haslett, J.N., Kunkel, L.M., et al., *Reproducibility of gene expression across generations of Affymetrix microarrays.* BMC Bioinformatics, 4: p. 27. 2003.
58. Autio, R., Kilpinen, S., Hautaniemi, S., and Kallioniemi, O., *Redefinition of probe sets improves the comparability of the data between Affymetrix array generations.* Proceedings of the 4th TICSP Workshop on Computational Systems Biology (WCSB 2006): p. 31-34. 2006.
59. Tillinghast, G.W., *Microarrays in the clinic.* Nat Biotechnol, 28(8): p. 810-2.
60. Jarvinen, A.K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., et al., *Are data from different gene expression microarray platforms comparable?* Genomics, 83(6): p. 1164-8. 2004.
61. Barnes, M., Freudenberg, J., Thompson, S., Aronow, B., and Pavlidis, P., *Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms.* Nucleic Acids Res, 33(18): p. 5914-23. 2005.
62. Glas, A.M., Floore, A., Delahaye, L.J., Witteveen, A.T., Pover, R.C., et al., *Converting a breast cancer microarray signature into a high-throughput diagnostic test.* BMC Genomics, 7: p. 278. 2006.
63. Affymetrix, *Statistical algorithms reference guide.* Technical Report, Affymetrix. 2001.
64. Li, C. and Hung Wong, W., *Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application.* Genome Biol, 2(8): p. RESEARCH0032. 2001.
65. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 4(2): p. 249-64. 2003.
66. Irizarry, R.A., Wu, Z., and Jaffee, H.A., *Comparison of Affymetrix GeneChip expression measures.* Bioinformatics, 22(7): p. 789-94. 2006.
67. Hubbard, T.J., Aken, B.L., Ayling, S., Ballester, B., Beal, K., et al., *Ensembl 2009.* Nucleic Acids Res, 37(Database issue): p. D690-7. 2009.
68. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., et al., *The human genome browser at UCSC.* Genome Res, 12(6): p. 996-1006. 2002.
69. Tatusova, T., *Genomic databases and resources at the National Center for Biotechnology Information.* Methods Mol Biol, 609: p. 17-44.
70. Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., et al., *Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.* Nucleic Acids Res, 33(20): p. e175. 2005.
71. Elo, L.L., Lahti, L., Skottman, H., Kylaniemi, M., Lahesmaa, R., et al., *Integrating probe-level expression changes across generations of Affymetrix arrays.* Nucleic Acids Research, 33(22): p. e193. 2005.
72. Mecham, B.H., Klus, G.T., Strovel, J., Augustus, M., Byrne, D., et al., *Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements.* Nucleic Acids Res, 32(9): p. e74. 2004.

73. Culhane, A.C., Perriere, G., and Higgins, D.G., *Cross-platform comparison and visualisation of gene expression data using co-inertia analysis.* BMC Bioinformatics, 4: p. 59. 2003.
74. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., et al., *ArrayExpress--a public repository for microarray gene expression data at the EBI.* Nucleic Acids Res, 31(1): p. 68-71. 2003.
75. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., et al., *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.* Nat Genet, 29(4): p. 365-71. 2001.
76. Ioannidis, J.P., Allison, D.B., Ball, C.A., Coulibaly, I., Cui, X., et al., *Repeatability of published microarray gene expression analyses.* Nat Genet, 41(2): p. 149-55. 2009.
77. Huminiecki, L., Lloyd, A.T., and Wolfe, K.H., *Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases.* BMC Genomics, 4(1): p. 31. 2003.
78. Lash, A.E., Tolstoshev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., et al., *SAGEmap: a public gene expression resource.* Genome Res, 10(7): p. 1051-60. 2000.
79. Skrabanek, L. and Campagne, F., *TissueInfo: high-throughput identification of tissue expression profiles and specificity.* Nucleic Acids Res, 29(21): p. E102-2. 2001.
80. Segal, E., Yelensky, R., and Koller, D., *Genome-wide discovery of transcriptional modules from DNA sequence and gene expression.* Bioinformatics, 19 Suppl 1: p. i273-82. 2003.
81. Xu, X., Wang, L., and Ding, D., *Learning module networks from genome-wide location and expression data.* FEBS Lett, 578(3): p. 297-304. 2004.
82. Segal, E., Friedman, N., Koller, D., and Regev, A., *A module map showing conditional activity of expression modules in cancer.* Nat Genet, 36(10): p. 1090-8. 2004.
83. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., et al., *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.* Nat Genet, 34(2): p. 166-76. 2003.
84. Cahan, P., Rovegno, F., Mooney, D., Newman, J.C., St Laurent, G., 3rd, et al., *Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization.* Gene, 401(1-2): p. 12-8. 2007.
85. Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., and Gruissem, W., *GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox.* Plant Physiol, 136(1): p. 2621-32. 2004.
86. Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., et al., *BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources.* Genome Biol, 10(11): p. R130. 2009.
87. Chen, X., Ji, Z.L., and Chen, Y.Z., *TTD: Therapeutic Target Database.* Nucleic Acids Res, 30(1): p. 412-5. 2002.
88. Greco, D., Somervuo, P., Di Lieto, A., Raitila, T., Nitsch, L., et al., *Physiology, pathology and relatedness of human tissues from gene expression meta-analysis.* PLoS One, 3(4): p. e1880. 2008.
89. Markowitz, V.M., *Data management challenges for molecular and cell biology: an industry perspective.* OMICS, 7(1): p. 121-2. 2003.
90. Uhlen, M., Bjorling, E., Agaton, C., Szigyarto, C.A., Amini, B., et al., *A human protein atlas for normal and cancer tissues based on antibody proteomics.* Mol Cell Proteomics, 4(12): p. 1920-32. 2005.
91. Leivonen, S.K., Makela, R., Ostling, P., Kohonen, P., Haapa-Paananen, S., et al., *Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines.* Oncogene, 28(44): p. 3926-36. 2009.
92. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., et al., *A census of human cancer genes.* Nat Rev Cancer, 4(3): p. 177-83. 2004.
93. Reese, D.M. and Slamon, D.J., *HER-2/neu signal transduction in human breast and ovarian cancer.* Stem Cells, 15(1): p. 1-8. 1997.
94. Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 98(19): p. 10869-74. 2001.
95. Miller, L.D., Smeds, J., George, J., Vega, V.B., Vergara, L., et al., *An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.* Proc Natl Acad Sci U S A, 102(38): p. 13550-5. 2005.
96. Pavey, S., Johansson, P., Packer, L., Taylor, J., Stark, M., et al., *Microarray expression profiling in melanoma reveals a BRAF mutation signature.* Oncogene, 23(23): p. 4060-7. 2004.
97. Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S., *The protein kinase complement of the human genome.* Science, 298(5600): p. 1912-34. 2002.
98. Milanesi, L., Petrillo, M., Sepe, L., Boccia, A., D'Agostino, N., et al., *Systematic analysis of human kinase genes: a large number of genes and alternative splicing events result in functional and structural diversity.* BMC Bioinformatics, 6 Suppl 4: p. S20. 2005.
99. Varjosalo, M., Bjorklund, M., Cheng, F., Syvanen, H., Kivioja, T., et al., *Application of active and kinase-deficient kinome collection for identification of kinases regulating hedgehog signaling.* Cell, 133(3): p. 537-48. 2008.
100. Johnson, S.A. and Hunter, T., *Kinomics: methods for deciphering the kinome.* Nat Methods, 2(1): p. 17-25. 2005.
101. Prifti, E., Zucker, J.D., Clement, K., and Henegar, C., *FunNet: an integrative tool for exploring transcriptional interactions.* Bioinformatics, 24(22): p. 2636-8. 2008.
102. Zhang, W., Morris, Q.D., Chang, R., Shai, O., Bakowski, M.A., et al., *The functional landscape of mouse gene expression.* J Biol, 3(5): p. 21. 2004.
103. Hu, P., Bader, G., Wigle, D.A., and Emili, A., *Computational prediction of cancer-gene function.* Nat Rev Cancer, 7(1): p. 23-34. 2007.
104. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 12(4): p. 656-64. 2002.
105. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., *Basic local alignment search tool.* J Mol Biol, 215(3): p. 403-10. 1990.
106. Parmigiani, G., Garrett, E.S., Anbazhagan, R., and Gabrielson, E., *A statistical framework for expression-based molecular classification in cancer.* Journal Of The Royal Statistical Society Series B, 64(4): p. 717-736. 2002.

107. Zilliox, M.J. and Irizarry, R.A., *A gene expression bar code for microarray data.* Nat Methods, 4(11): p. 911-3. 2007.
108. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., et al., *The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.* Science, 313(5795): p. 1929-35. 2006.
109. Caldas, J., Gehlenborg, N., Faisal, A., Brazma, A., and Kaski, S., *Probabilistic retrieval and visualization of biologically relevant microarray experiments.* Bioinformatics, 25(12): p. i145-53. 2009.
110. Lopez, F., Textoris, J., Bergon, A., Didier, G., Remy, E., et al., *TranscriptomeBrowser: a powerful and flexible toolbox to explore productively the transcriptional landscape of the Gene Expression Omnibus database.* PLoS One, 3(12): p. e4001. 2008.
111. Gruvberger-Saal, S.K., Cunliffe, H.E., Carr, K.M., and Hedenfalk, I.A., *Microarrays in breast cancer research and clinical practice--the future lies ahead.* Endocr Relat Cancer, 13(4): p. 1017-31. 2006.
112. Roepman, P., Horlings, H.M., Krijgsman, O., Kok, M., Bueno-de-Mesquita, J.M., et al., *Microarray-based determination of estrogen receptor, progesterone receptor, and HER2 receptor status in breast cancer.* Clin Cancer Res, 15(22): p. 7003-11. 2009.
113. Sarwate, D.V., *Computation of Cyclic Redundancy Checks Via Table Lookup.* Communications of the ACM,(August). 1988.
114. Hautaniemi, S., Kauraniemi, P., Rämö, P., Yli-Harja, O., Astola, J., Kallioniemi, A. . *A strategy for identifying class-separating genes in drug-treatment microarray data, Report 1.* 2003. Institute of Signal Processing, Tampere University of Technology, Finland.
115. Khan, J., Simon, R., Bittner, M., Chen, Y., Leighton, S.B., et al., *Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays.* Cancer Res, 58(22): p. 5009-13. 1998.
116. Hubert Lawrence, A.P., *Comparing partitions.* Journal of classification,(2): p. 193-218. 1985.
117. Kullback, S., *Letter to the Editor: The Kullback–Leibler distance.* The American Statistician, 4(41): p. 340-341. 1987.
118. Kullback, S. and Leibler, R.A., *On Information and Sufficiency.* Annals of Mathematical Statistics,(22): p. 79-86. 1951.
119. Frohlich, H., Speer, N., Poustka, A., and Beissbarth, T., *GOSim--an R-package for computation of information theoretic GO similarities between terms and gene products.* BMC Bioinformatics, 8: p. 166. 2007.
120. Scott, D.W. and Härdle, W., *Smoothing by weighted averaging of rounded points.* Computational Statistics,(7:97). 1992.
121. Lin, D., *An information-theoretic definition of similarity.* Proceedings of the 15th International Conference on Machine Learning, 1: p. 296-304. 1998.
122. Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., et al., *Ensembl 2008.* Nucleic Acids Res, 36(Database issue): p. D707-14. 2008.
123. Yeoh, E.J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., et al., *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.* Cancer Cell, 1(2): p. 133-43. 2002.
124. Ross, M.E., Zhou, X., Song, G., Shurtleff, S.A., Girtman, K., et al., *Classification of pediatric acute lymphoblastic leukemia by gene expression profiling.* Blood, 102(8): p. 2951-9. 2003.
125. Isotalo, P.A., Greenway, D.C., and Donnelly, J.G., *Metastatic alveolar rhabdomyosarcoma with increased serum creatine kinase MB and cardiac troponin T and normal cardiac troponin I.* Clin Chem, 45(9): p. 1576-8. 1999.
126. Plouzek, C.A., Leslie, K.K., Stephens, J.K., and Chou, J.Y., *Differential gene expression in the amnion, chorion, and trophoblast of the human placenta.* Placenta, 14(3): p. 277-85. 1993.
127. Kellen, J.A., Bush, R.S., and Malkin, A., *Placenta-like alkaline phosphatase in gynecological cancers.* Cancer Res, 36(1): p. 269-71. 1976.
128. Ind, T.E., Iles, R.K., Carter, P.G., Lowe, D.G., Shepherd, J.H., et al., *Serum placental-type alkaline phosphatase activity in women with squamous and glandular malignancies of the reproductive tract.* J Clin Pathol, 47(11): p. 1035-7. 1994.
129. Philippe, E., Omlin, F.X., and Droz, B., *Myelin-associated glycoprotein immunoreactive material: an early neuronal marker of dorsal root ganglion cells during chick development.* Brain Res, 392(1-2): p. 275-7. 1986.
130. Shaw, J.L. and Diamandis, E.P., *Distribution of 15 human kallikreins in tissues and biological fluids.* Clin Chem, 53(8): p. 1423-32. 2007.
131. Brenner, M., Kisseberth, W.C., Su, Y., Besnard, F., and Messing, A., *GFAP promoter directs astrocyte-specific expression in transgenic mice.* J Neurosci, 14(3 Pt 1): p. 1030-7. 1994.
132. Kalejs, M. and Erenpreisa, J., *Cancer/testis antigens and gametogenesis: a review and "brain-storming" session.* Cancer Cell Int, 5(1): p. 4. 2005.
133. Epping, M.T., Wang, L., Edel, M.J., Carlee, L., Hernandez, M., et al., *The human tumor antigen PRAME is a dominant repressor of retinoic acid receptor signaling.* Cell, 122(6): p. 835-47. 2005.
134. Miettinen, M. and Lasota, J., *Gastrointestinal stromal tumors: review on morphology, molecular pathology, prognosis, and differential diagnosis.* Arch Pathol Lab Med, 130(10): p. 1466-78. 2006.
135. Zhu, J., He, F., Song, S., Wang, J., and Yu, J., *How many human genes can be defined as housekeeping with current expression data?* BMC Genomics, 9: p. 172. 2008.
136. Paradis, E., Claude, J., and Strimmer, K., *APE: Analyses of Phylogenetics and Evolution in R language.* Bioinformatics, 20(2): p. 289-90. 2004.
137. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 25(1): p. 25-9. 2000.
138. Ohishi, T., Hirota, T., Tsuruo, T., and Seimiya, H., *TRF1 mediates mitotic abnormalities induced by Aurora-A overexpression.* Cancer Res, 70(5): p. 2041-52.
139. Klebig, C., Korinth, D., and Meraldi, P., *Bub1 regulates chromosome segregation in a kinetochore-independent manner.* J Cell Biol, 185(5): p. 841-58. 2009.
140. Brennan, I.M., Peters, U., Kapoor, T.M., and Straight, A.F., *Polo-like kinase controls vertebrate spindle elongation and cytokinesis.* PLoS One, 2(5): p. e409. 2007.

141. Huang, Y.F., Chang, M.D., and Shieh, S.Y., *TTK/hMps1 mediates the p53-dependent postmitotic checkpoint by phosphorylating p53 at Thr18.* Mol Cell Biol, 29(11): p. 2935-44. 2009.
142. Gavet, O. and Pines, J., *Progressive activation of CyclinB1-Cdk1 coordinates entry to mitosis.* Dev Cell, 18(4): p. 533-43.
143. Gaudet, S., Branton, D., and Lue, R.A., *Characterization of PDZ-binding kinase, a mitotic kinase.* Proc Natl Acad Sci U S A, 97(10): p. 5167-72. 2000.
144. Chan, G.K., Jablonski, S.A., Sudakin, V., Hittle, J.C., and Yen, T.J., *Human BUBR1 is a mitotic checkpoint kinase that monitors CENP-E functions at kinetochores and binds the cyclosome/APC.* J Cell Biol, 146(5): p. 941-54. 1999.
145. Habedanck, R., Stierhof, Y.D., Wilkinson, C.J., and Nigg, E.A., *The Polo kinase Plk4 functions in centriole duplication.* Nat Cell Biol, 7(11): p. 1140-6. 2005.
146. Fry, A.M., Meraldi, P., and Nigg, E.A., *A centrosomal function for the human Nek2 protein kinase, a member of the NIMA family of cell cycle regulators.* Embo J, 17(2): p. 470-81. 1998.
147. Zhang, Y.W., Brognard, J., Coughlin, C., You, Z., Dolled-Filhart, M., et al., *The F box protein Fbx6 regulates Chk1 stability and cellular sensitivity to replication stress.* Mol Cell, 35(4): p. 442-53. 2009.
148. Chan, Y.W., Fava, L.L., Uldschmid, A., Schmitz, M.H., Gerlich, D.W., et al., *Mitotic control of kinetochore-associated dynein and spindle orientation by human Spindly.* J Cell Biol, 185(5): p. 859-74. 2009.
149. Molinaro, A.M., Simon, R., and Pfeiffer, R.M., *Prediction error estimation: a comparison of resampling methods.* Bioinformatics, 21(15): p. 3301-7. 2005.
150. Duda, R.O. and Hart, P.E., *Nonparametric Techniques.* In Pattern Classification and Scene Analysis: p. 98-105. 1973.
151. Fukunaga, K., *Nonparametric Classification and Error Estimation.* In Introduction to statistical pattern recognition: p. 303-322. 1990.
152. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., et al., *Multiclass cancer diagnosis using tumor gene expression signatures.* Proc Natl Acad Sci U S A, 98(26): p. 15149-54. 2001.
153. Yeang, C.H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R.M., et al., *Molecular classification of multiple tumor types.* Bioinformatics, 17 Suppl 1: p. S316-22. 2001.
154. Mjolsness, E. and DeCoste, D., *Machine learning for science: state of the art and future prospects.* Science, 293(5537): p. 2051-5. 2001.
155. Forner, F., Kumar, C., Luber, C.A., Fromme, T., Klingenspor, M., et al., *Proteome differences between brown and white fat mitochondria reveal specialized metabolic functions.* Cell Metab, 10(4): p. 324-35. 2009.
156. Hu, E., Liang, P., and Spiegelman, B.M., *AdipoQ is a novel adipose-specific gene dysregulated in obesity.* J Biol Chem, 271(18): p. 10697-703. 1996.
157. Urs, S., Smith, C., Campbell, B., Saxton, A.M., Taylor, J., et al., *Gene expression profiling in human preadipocytes and adipocytes by microarray analysis.* J Nutr, 134(4): p. 762-70. 2004.
158. Zhu, Q., Anderson, G.W., Mucha, G.T., Parks, E.J., Metkowski, J.K., et al., *The Spot 14 protein is required for de novo lipid synthesis in the lactating mammary gland.* Endocrinology, 146(8): p. 3343-50. 2005.
159. Shmulevich, I. and Zhang, W., *Binary analysis and optimization-based normalization of gene expression data.* Bioinformatics, 18(4): p. 555-65. 2002.
160. Pal, R., Datta, A., Fornace, A.J., Jr., Bittner, M.L., and Dougherty, E.R., *Boolean relationships among genes responsive to ionizing radiation in the NCI 60 ACDS.* Bioinformatics, 21(8): p. 1542-9. 2005.